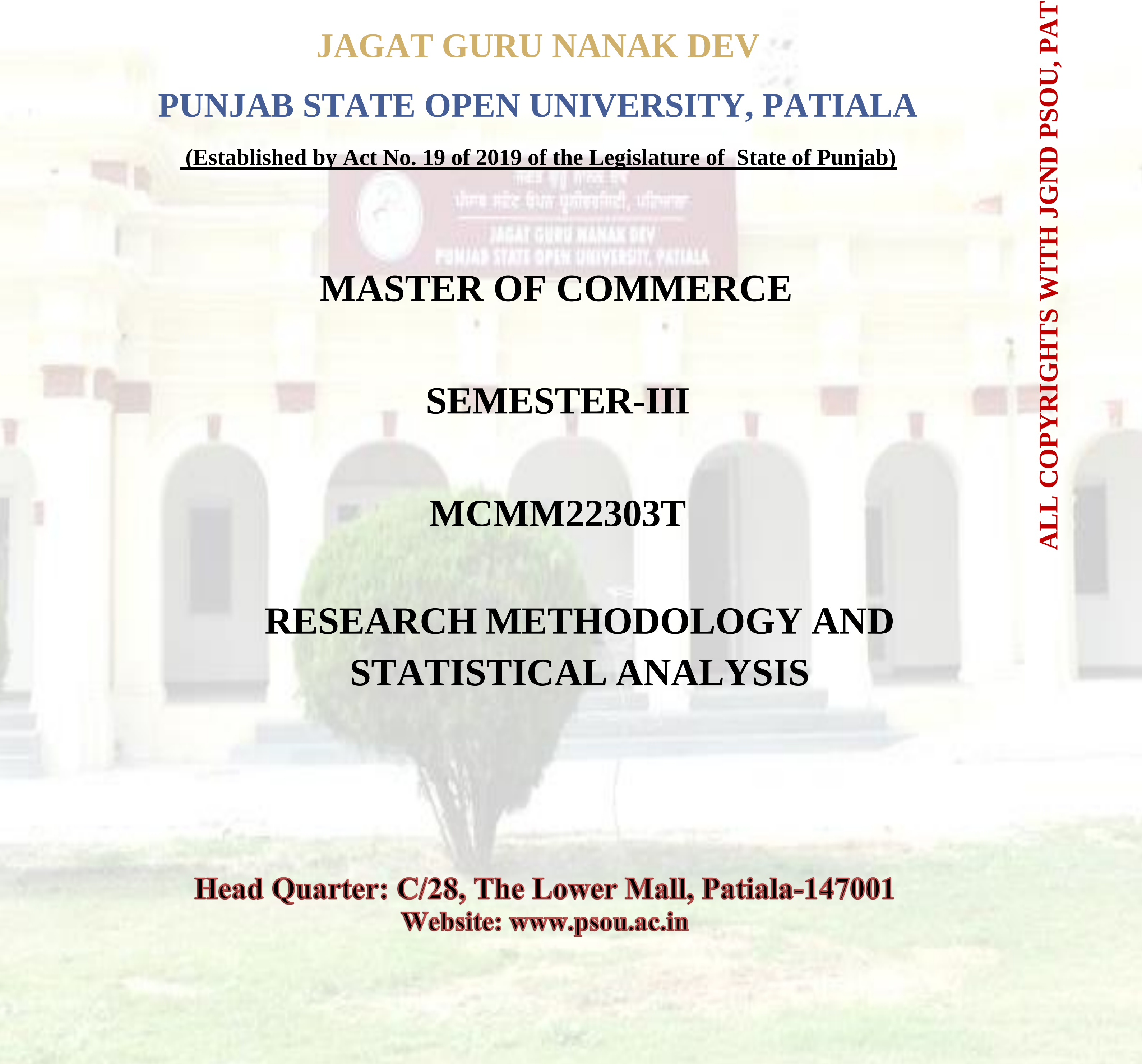
ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

The Motto of Our University
(SEWA)
SKILL ENHANCEMENT
EMPLOYABILITY
WISDOM
ACCESSIBILITY



JAGAT GURU NANAK DEV

PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

MASTER OF COMMERCE

SEMESTER-III

MCMM22303T

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

Head Quarter: C/28, The Lower Mall, Patiala-147001
Website: www.psou.ac.in

The Study Material has been prepared exclusively under the guidance of Jagat Guru Nanak Dev Punjab State Open University, Patiala, as per the syllabi prepared by Committee of experts and approved by the Academic Council.

COURSE COORDINATOR AND EDITOR:

Dr. Pinky Sra

Assistant Professor

JGND PSOU, Patiala.

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

JAGAT GURU NANAK DEV PUNJAB STATE OPEN UNIVERSITY,PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

PREFACE

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in December 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open University of the State, entrusted with the responsibility of making higher education accessible to all, especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self- instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The University has a network of 110 Learner Support Centres/Study Centres, to enable students to make use of reading facilities, and for curriculum-based counseling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G.S. Batra
Dean Academic Affairs

MASTER OF COMMERCE (M.COM)

SEMESTER-III

(MCMM22303T): RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

MAX. MARKS: 100

EXTERNAL: 70

INTERNAL: 30

PASS: 40%

CREDITS: 6

Objective: To enable students to identify various data collection methods for the purpose of research and the statistical tools required for the analysis of data.

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER:

1. The syllabus prescribed should be strictly adhered to.
2. The question paper will consist of three sections: A, B, and C. Sections A and B will have four questions from the respective sections of the syllabus and will carry 10 marks each. The candidates will attempt two questions from each section.
3. Section C will have fifteen short answer questions covering the entire syllabus. Each question will carry 3 marks. Candidates will attempt any ten questions from this section.
4. The examiner shall give a clear instruction to the candidates to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.
5. The duration of each paper will be three hours.

INSTRUCTIONS FOR THE CANDIDATES:

Candidates are required to attempt any two questions each from sections A and B ofthe question paper and any ten short questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.

SECTION-A

Unit 1: Research and Data Collection Introduction to Business Research, Research Plan, Collection of Data

Unit 2: Sample Measurement and Scaling Techniques

Unit 3: Processing and Preservation of Data: Processing of Data, Diagrammatic and Graphic Presentation

Unit 4: Measures of Central Tendency

Unit 5: Measures of Variation and Skewness

SECTION-B

Unit 6: Relational and Trend Analysis: Correlation and Simple Regression

Unit 7: Time Series Analysis and Index Numbers

Unit 8: Probability and Hypothesis Testing: Probability and Probability Rules Probability Distributions

Unit 9: Tests of Hypothesis–I Tests of Hypothesis – II, Chi-Square Test

Unit 10: Interpretation and Reporting: Interpretation of Statistical DataReport Writing

Suggested Readings:

- Cooper, D. R., and Schindler, P.S., "Business Research Methods", 9th Edition, Tata McGraw-Hill, New Delhi.
- Levine, D.M., Krehbiel T.C. and Berenson M.L., "Business Statistics", 12th Edition (2012), Pearson Education, New Delhi.
- Kothari, C. R., "Research Methodology", 2nd Edition (2008), New Age International.
- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and Economics", 2nd edition (2011), Thompson, New Delhi.
- http://swayam.gov.in/

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

MASTER OF COMMERCE (M.COM)

SEMESTER-III

(MCMM22303T): RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

EDITOR AND COURSE CO-ORDINATOR- DR. PINKY SRA

SECTION A

UNIT NO.	UNIT NAME
Unit 1	Research and Data Collection Introduction to Business Research, Research Plan, Collection of Data
Unit 2	Sample Measurement and Scaling Techniques
Unit 3	Processing and Preservation of Data: Processing of Data, Diagrammatic and Graphic Presentation
Unit 4	Measures of Central Tendency
Unit 5	Relational and Trend Analysis: Correlation and Simple Regression

SECTION B

UNIT NO.	UNIT NAME
Unit 6	Measures of Central Tendency- Mean (Direct, Short cut and step deviation methods), Merits & Demerits.
Unit 7	Time Series Analysis and Index Numbers
Unit 8	Probability and Hypothesis Testing: Probability and Probability Rules Probability Distributions
Unit 9	Tests of Hypothesis–I Tests of Hypothesis – II, Chi-Square Test
Unit 10	Interpretation and Reporting: Interpretation of Statistical Data Report Writing

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 1: RESEARCH AND DATA COLLECTION INTRODUCTION TO BUSINESS RESEARCH, RESEARCH PLAN, COLLECTION OF DATA

STRUCTURE

1.0 LEARNING OBJECTIVES

After reading the unit, learners will be able to learn:

- Meaning of business research
- Research plan
- Types of data collection
- Meaning of Primary data and Secondary data
- Limitations of Primary and Secondary data
- Precautions to Collect Secondary Data

1.1 INTRODUCTION OF BUSINESS RESEARCH

Business research is the process of gathering thorough data and information from all business areas and applying this information to increase sales and profit. In order to gain more knowledge and make wise business decisions, business research refers to the investigation and analysis of numerous business-related issues. Business Research includes collection of the data, interpretation, and evaluation of data and information related to markets, customers, competitors, and other relevant factors which influence business operations.

The main objective of business research is to obtain the knowledge and understanding that can be applied to improve business performance, solve problems, identify opportunities, and support decision-making processes. Businesses can use research to collect insightful data and information that can help them create winning strategies, streamline processes, and improve their competitiveness. Business research also includes a wide range of areas including market research, consumer behavior analysis, product development, pricing strategies, marketing effectiveness, operational efficiency, financial analysis, and more. It can be conducted through various methods, such as surveys, interviews, observations, experiments, and data analysis.

For businesses of all sizes and in every industry, business research is essential. Businesses may make smart judgments and adjust to shifting market conditions by staying updated about market trends, client preferences, and competitive landscapes. Businesses can reduce risks, spot chances for innovation and growth, and get a better understanding of their target market by conducting research, all of which can help them succeed in the long run.

1.2 ROLE OF BUSINESS RESEARCH

Business research plays a crucial role in assisting organizations in making wise decisions, resolving issues, and achieving their strategic goals. Here are some key roles that business research plays:

- The main role of business research is to help across every decision in the business, starting from product innovation to marketing and promotional planning.
- Business research helps in foreseeing a company's future challenges, including those related to competition.
- New product development and innovation are greatly facilitated by research. Before allocating resources to the development, businesses can find unmet requirements, gather ideas for new features or products, and validate concepts by doing market research, customer surveys, and analysis of consumer feedback.
- Ensuring customer happiness is another important area where this has a greater impact since through research, we can find areas where we can effectively serve our target audience.
- Business research can help an organisation implement cost-effectiveness by helping to minimize costs where they are necessary and spend more money when profit is generated.
- Businesses may remain ahead of the competition by conducting business research. Businesses can develop competitive strategies, set themselves apart from competitors, and identify opportunities for acquiring a competitive edge in the market by analyzing the strengths and weaknesses, market positioning, and strategic movements of competitors.

1.3 ADVANTAGES OF BUSINESS RESEARCH

- Business research used to recognize potential risks, problems, and opportunities.
- To offer a customer and target audience study, so assisting in bettering relationships with one's audience and capturing the areas which we might be missing out on.
- Businesses can use business research to locate and evaluate competitors as well as their advantages and disadvantages in the market. This information enables companies to create competitive strategies, set themselves apart from competitors, and take advantage of market opportunities to acquire a competitive edge.
- To predict future issues so that the business can deal with such concerns

- It provides firms with the opportunity to develop stronger plans to compete with their competitors by continuously tracking market competition.
- Business Research also does a complete cost study, assisting the organisation in resource management and allocation.
- It updates you on the most recent trends and competitive analysis.

1.4 DISADVANTAGES OF BUSINESS RESEARCH

- It can be time- and money-consuming.
- Bias can influence the research process and interpretation of findings. Bias may be by mistake introduced by researchers when gathering, analysing, or reporting data.
- The focus groups may be small or heavy, based on assumptions, which increases the risk of it being at times assuming and inaccurate.
- The quality and availability of data can present challenges in business research. Depending on the extent of the research, some data may be unavailable or partial, which could have an impact on the analysis and conclusions. It could take a lot of time and money to get accurate and current data.
- Because of the market's constant evolution and change, it might be difficult for business research to identify or predict the proper trends.

1.5 RESEARCH PLAN

A business research plan includes a structured approach to identify research objectives, determine the research methodology, plan the data collection process, and method of data collection and establish a timeline and budget. Here is a framework to help you create a business research plan:

1. **Research Objectives**: State the purpose of the research and the specific objectives you want to achieve. What issue or query are you attempting to solve? What facts or insights are you looking for?
2. **Literature Review**: Review existing literature, research studies, and industry reports related to your research topic. This step helps you understand the existing knowledge and identify any research gaps that need to be addressed.
3. **Research Methodology**: To identify the research methodology based on research objectives.

Commonly surveys, interviews, observations, experiments, focus groups, or a combination of these are included in the method of data collection. Justify why you chose the specific methodology and how it aligns with your objectives.

4. **Sample Selection**: To determine the total population as well as sample size for your research. Consider the characteristics of the population you want to study and define the sampling technique to ensure it represents your target audience very accurately.
5. **Data Collection Tools**: Create the appropriate data gathering tools, such as surveys, interview guides, or checklists for observation. Just make sure the tools are accurate, trustworthy, and compatible with your research goals.
6. **Data Collection:** Make a plan for the data collection procedure. How to talk to people, plan surveys or interviews, or make observations. Think about participant permission, data protection, and ethical issues.
7. **Data Analysis**: Define the methods or techniques to examine the collected data. This basically, involves either qualitative analysis (e.g., thematic analysis, content analysis) or quantitative analysis (e.g., statistical analysis, data modeling). Also, the choice of appropriate software or tools for data analysis is important.
8. **Interpretation**: here we consider how to interpret the research findings. Identify key themes, patterns, or trends in the data and relate them back to your research objectives and draw meaningful conclusions on the basis of an analysis.
9. **Policy Implications**: these must be based on research findings, and propose real-world policy as well as strategies that covered the research objectives. These recommendations should provide actionable insights to guide decision-making and drive business improvements.
10. **Representation of Results**: This may include preparing reports, presentations, or visualizations to effectively communicate the results to relevant stakeholders.
11. **Create a Timeline and Budget**: Budget and allocate resources to pay for expenditures associated with data collection, analysis, participant incentives, and other research-related expenses.

In the research plan, it is to make sure it remains in accordance with your objectives and takes into account any new factors or difficulties, keep reviewing and revising your research plan as necessary during the study process.

1.6 COLLECTION OF DATA: MEANING

The most important part of conducting business research is data collection. In order to fulfil research objectives and respond to research questions, it involves gathering relevant data and insights. Data collection refers to the systematic process of gathering information and data to address research objectives or answer specific research questions. It involves collecting relevant facts, figures, observations, or responses from various sources or participants, using specific methods or instruments. The purpose of data collection is to obtain correct, consistent, and meaningful data that can be analysed as well as interpreted to draw insights and make informed decisions. It is a critical step in the research process as the quality of the data collected directly impacts the validity and reliability of the research findings.

Several important factors for data collection in business research are like Identify Data Sources, Determine Sample Size and Selection, Design Data Collection Instruments, Pilot Testing, Data Collection Procedures, Ensure Data Quality, Record and Organize Data, and Data Verification and Cleaning.

1.7 TYPES OF DATA COLLECTION

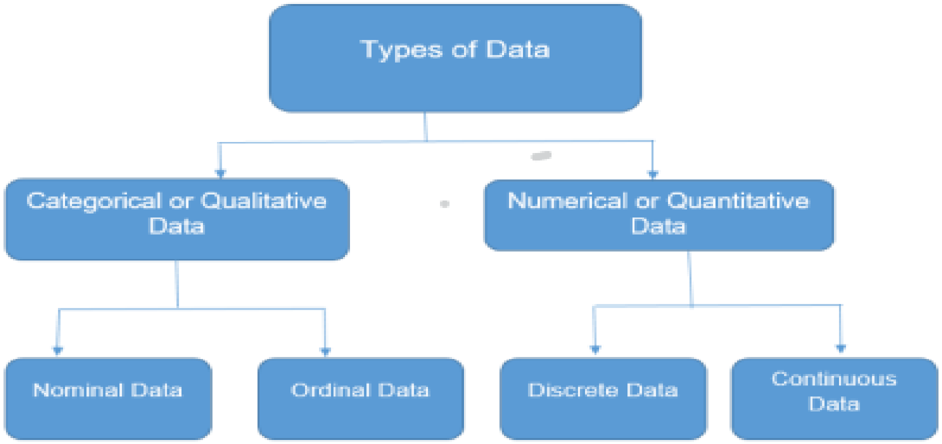
By now you have known that data could be classified in the following three ways:

a) Quantitative and Qualitative Data

b) Sample and Census Data

c) Primary and Secondary data

1.7.1 Quantitative and Qualitative data: Quantitative data are those set of information that are quantifiable and can be expressed in some standard units like rupees, kilograms, litters, etc. For example, pocket money of students of a class and the income of their parents can be expressed in so many rupees; the production or import of wheat can be expressed in so many kilograms or lakh quintals; the consumption of petrol and diesel in India as so many lakh litters in one year and so on. In other words, Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are de



fined in terms of natural language specifications, but not in terms of numbers

1. **Qualitative Data:** on the other hand, are not quantifiable, that is, cannot be expressed in standard units of measurement like rupees, kilograms, liters, etc. This is because they are 'features', 'qualities', or 'characteristics' like eye colors, skin complexion, honesty, good or bad, etc. These are also referred to as attributes. In this case, however, it is possible to count the number of individuals (or items) possessing a particular attribute.

- **Nominal Data:** Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc. The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.
- **Ordinal Data**: Ordinal data is a type of data which follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on. The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualization tools. The information may be expressed using tables in which each row in the table shows the distinct category. Quantitative or Numerical Data.

2. **Quantitative Data:** Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information

about the quantities of aspecific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. Thetwo different classifications of numerical data are discrete data and continuous data.

- **Discrete Data:** Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in the whole numbers e.g., Number of students in the class
- **Continuous Data:** Continuous data is data that can be calculated. It has an infinite number ofprobable values that can be selected within a given specific range e.g., Temperature range.

1.7.2 Sample and Census Data: Data can be collected either by census method or sample method. Information collected through sample inquiry is called sample data and the one collected through census inquiry is called census data. Population census data are collected every ten years in India.

1.7.3 Primary and Secondary Data: Primary data are collected by the investigator through field surveys. Such data are in raw form and must be refined before use. On the other hand, secondary data are extracted from the existing published or unpublished sources, that is; from the data already collected by others. The collection of data is the first basic step towards the statistical analysis of any problem. The collected data are suitably transformed and analysed to draw conclusions about the population.

These conclusions may be either or both of the following:

i) To estimate one or more parameters of a population or the nature of the population itself. This forms the subject matter of the theory of estimation.

ii) To test a hypothesis. A hypothesis is a statement regarding the parameters or the nature of the population.

1.8 SOURCES OF DATA COLLECTION

A pertinent question that arises now is how and from where to get data? Data are obtained through two types of investigations, namely,

1) Direct Investigation **or Primary Data** which implies that the investigator collects information by observing the items of the problem under investigation. As explained above, it is the primary

source of getting data or the source of getting primary data and can be done through observation or through inquiry. In the former we watch an event happening, for example, the number and type of vehicles passing through Vijay Chowk in New Delhi during different hours of the day and night. In the latter, we ask questions from the respondents through questionnaire (personally or through mail). It is a costly method in terms of money, time, and effort.

2) Investigation through **Secondary Source** which means obtaining data from the already collected data. Secondary data are the other people's statistics, where other people include governments at all levels, international bodies or institutions like IMF, IBRD, etc., or other countries, private and government research organisations, Reserve Bank of India and other banks, research scholars of repute, etc. Broadly speaking we can divide the sources of secondary data into two categories: published sources and unpublished sources. A) Published Sources

- Official publications of the government at all levels — Central, State, Union
- Official publications of foreign countries.
- Official publications of international bodies like IMF, UNESCO, WHO, etc.
- Newspapers and Journals of repute, both local and international.
- Official publications of RBI, and other Banks, LIC, Trade Unions, Stock Exchange, Chambers of Commerce, etc.
- Reports submitted by reputed economists, research scholars, universities, commissions of inquiry, if made public.

Data Collection Methods Some main sources of published data in India are: Central Statistical Organisation (C.S.0.): It publishes data on national income, savings, capital formation, etc. in a publication called National Accounts Statistics. National Sample Survey Organisation (N.S.S.O.): Under Ministry of Statistics and Programme Implementation, this organisation provides us data on all aspects of national economy, such as agriculture, industry, labour and consumption expenditure. Reserve Bank of India Publications (R.B.I.): It publishes financial statistics. Its publications are Report on Currency and Finance, Reserve Bank of India Bulletin, Statistical Tables Relating to Banks in India, etc. iv) Labour Bureau: Its publications are Indian Labour Statistics, Indian Labour Year Book, Indian Labour Journal, etc. v) Population Census: Undertaken by the office of the Registrar General India, Ministry of Home Affairs. It provides us different types of statistics about population.

B) Un-published Sources

- Unpublished findings of certain inquiry committees.
- Research workers' findings.
- Unpublished material found with Trade Associations, Labour Organisations and Chambers of Commerce.

CHECK YOUR PROGRESS (A)

Q1. Explain the term quantitative data.

Ans:____

____

Q2. What is primary data?

Ans:____

____

Q3. What is secondary data set?

Ans:____

____

Q4. Define sample and population.

Ans:____

____

1.9 COLLECTION OF PRIMARY DATA SURVEY TECHNIQUES

After the investigator is convinced that the gain from primary data outweighs the money cost, effort and time, she/he can go in for this. She/he can use any of the following methods to collect primary data:

a. Direct Personal Investigation
b. Indirect Oral Investigation
c. Use of Local Reports/ agencies to get information
d. Mailed Questionnaire Method
e. Schedules sent through enumerators

a) **Direct Personal Investigation**: Here the investigator collects information personally from the respondents. She/ he meets them personally to collect information. This method requires much from the investigator such as:

- She/he should be polite, unbiased and tactful.
- She/he should know the local conditions, customs and traditions
- She/he should be intelligent possessing good observation power. Data Collection Methods
- She/he should use simple, easy and meaningful questions to extract information.

This method is suitable only for intensive investigations. It is a costly method in terms of money, effort and time. Further, the personal bias of the investigator cannot be ruled out and it can do a lot of harm to the investigation. The method is a complete flop if the investigator does not possess the above-mentioned qualities.

b) Indirect Oral Investigation: Method This method is generally used when the respondents are reluctant to part with the information due to various reasons. Here, the information is collected from a witness or from a third party who are directly or indirectly related to the problem and possess sufficient knowledge. The person(s) who is/are selected as informants must possess the following qualities:

- They should possess full knowledge about the issue.
- They must be willing to reveal it faithfully and honestly.
- They should not be biased and prejudiced.
- They must be capable of expressing themselves to the true spirit of the inquiry.

c) Use of Local Reports: This method involves the use of local newspaper, magazines and journals by the investigators. The information is collected by local press correspondents and not by the investigators. Needless to say, this method does not yield sufficient and reliable data The method is less costly but should not be adopted where high degree of accuracy or precision is required.

d) Mailed Questionnaire Method: It is the most important and systematic method of collecting primary data, especially when the inquiry is quite extensive. This method entails creating a questionnaire (a collection of questions pertaining to the research area with a chance for respondents to fill in their replies) and mailing it to the respondents with a deadline for responding quickly. the respondents are asked to extend their full cooperation by providing accurate responses and timely submission of the completed questionnaire. By assuring them that the information they provided in the questionnaire will be kept totally secure and hidden, respondents are also given a sense of security. The investigator typically pays the return postal costs by mailing a self-addressed, stamped envelope to achieve a speedy and better response. Researchers, individuals, non-governmental organisations, and occasionally even the government involved in this technique.

e) Schedules sent through Enumerators: Using enumerators for primary data collection is a common practice in various research studies, surveys, and data collection efforts. Enumerators are individuals responsible for collecting data directly from respondents in the field. this is the method of obtaining answers to the questions in a form that is filled out by the interviewers or enumerators (the field agents who put these questions) in a face-to-face situation with the respondents. Questionnaire is a list of questions that the respondent himself answers in his own handwriting. 'Schedules sent through the enumerators' is the major data collecting technique that is commonly used. This is the case because the earlier ways that have been explained thus far have some drawbacks that this method does not. With the schedule (a list of questions), the enumerators directly contact the respondents, ask them the questions, and record their responses.

The questionnaire in primary data is divided into two parts:

1) General introductory part which contains questions regarding the identity of the respondent and contains information such as name, address, telephone number, qualification, profession, etc.
2) Main question part containing questions connected with the inquiry. These questions differ from inquiry-to-inquiry Preparation of the questionnaire is a highly specialized job and is perfected with experience. Therefore, some experienced persons should be associated with it.

Drafting and framing a questionnaire is a critical step in primary data collection. A well-designed questionnaire ensures that you gather relevant and reliable data to address your research objectives. The following few important points should be kept in mind while drafting a questionnaire:

(i) Clearly outline the research objectives and the specific information you want to collect through the questionnaire. Identify the key research questions that need to be answered.
(ii) Make sure your questions are easy to understand. Avoid nonsense and complex language. Keep sentences and questions short and to the point.

(iii) The task of soliciting information from people in desired form and with sufficient accuracy is the most difficult problem. By their nature people are not willing to reveal any information because of certain fears. Many a times they provide incomplete and faulty information. Therefore, it is necessary that the respondents be taken into confidence. They should be assured that their individual information will be kept confidential and no part of it will be revealed to tax and other government investigative agencies. This is very essential indeed. Where providing information is not legally binding, the informant has to be sure and convinced that the results of the survey will help the authorities to frame policies which will ultimately benefit them. It is obvious that some element of good salesmanship is also required in the investigation.

(iv) Make a decision regarding the questions that will be included in the questionnaire. Typical sorts of queries include:

- Closed-ended inquiries: Those who respond select from a set of predetermined responses (such as multiple-choice inquiries).
- Open-ended inquiries: Those that respond provide their own, individual responses.
- Questions using a likert scale: Determine if respondents agree or disagree with a statement using a scale (such as 1 to 5).
- Semantic differential questions: Request a rating on a scale of good to bad or satisfied to dissatisfied from respondents.

(v) Questions hurting the sentiments of respondent should not be asked. These include questions on his gambling habits, sex habits, indebtedness, etc.

(vi) Questions involving lengthy and complex calculations should be avoided because they require tedious extra work in which the respondent may lack both interests as well as capabilities.

1.10 LIMITATIONS OF PRIMARY DATA COLLECTION

Primary data refers to data collected firsthand through direct observation, surveys, interviews, experiments, or other data collection methods. While primary data can be valuable for research and analysis, it also has certain limitations. Here are some common limitations of primary data:

1. **Cost and time**: Collecting primary data can be a time-consuming and costly process. It requires resources to design research instruments, recruit participants, conduct data collection, and analyze the data. Therefore, primary data collection may be impractical or unaffordable.

2. **Limited sample size:** Primary data collection often involves a smaller sample size compared to secondary data sources. The sample size may be constrained by factors such as time, budget, or accessibility of the target population. A small sample size may limit the generalizability of the findings to a larger population.
3. **Sampling bias**: Similar to the limitations of statistics, primary data collection can be susceptible to sampling bias. If the sample is not representative of the population of interest, the findings may not accurately reflect the characteristics or behaviors of the larger population. Careful attention must be given to sampling methods to minimize bias.
4. **Response bias**: Response bias occurs when participants in a study provide inaccurate or misleading responses. It can be influenced by factors such as social desirability bias (participants providing responses they think are socially acceptable) or recall bias (participants inaccurately remembering past events). Response bias can undermine the validity and reliability of primary data.
5. **Subjectivity and researcher bias**: Primary data collection methods often involve interaction between the researcher and participants. The subjective interpretation and biases of the researcher can unintentionally influence the data collection process and the responses obtained. Researchers need to be aware of their own biases and take steps to minimize their impact on the data.
6. **Limited scope**: Primary data collection typically focuses on specific research questions or objectives. While this targeted approach can yield detailed insights into specific areas of interest, it may not capture a broader range of factors or provide a comprehensive understanding of the phenomenon being studied. Using secondary data or employing a mixed-methods approach can help overcome this limitation.
7. **Ethical considerations:** Primary data collection involves ethical considerations regarding participant privacy, informed consent, and data protection. Researchers must adhere to ethical guidelines and obtain necessary approvals, which can introduce additional time and logistical constraints.

Understanding these limitations of primary data can help researchers and analysts make informed decisions about data collection methods and consider the strengths and weaknesses of primary data in relation to their research objectives. It may also be beneficial to supplement primary data with secondary data sources to enhance the breadth and depth of the analysis.

CHECK YOUR PROGRESS (B)

Q1. Explain direct personal investigation and indirect oral investigation

Ans.____

____

Q2. Define Mailed questionnaire method and schedules sent through enumerators

Ans.____

____

Q3. Give limitations of primary data

Ans.____

____

1.11 COLLECTION OF SECONDARY DATA SOURCES

As direct investigation, though desirable, is costly in terms of money, time and efforts. Alternatively, information can also be obtained through a secondary source. It means drawing or collecting data from the already collected data of some other agency. Technically, the data so collected are called secondary data.

Secondary data sources in statistics refer to existing data that has been collected by someone else or for a different purpose but can be utilized for statistical analysis. These sources provide a wealth of information that can be used to explore research questions, test hypotheses, and derive insights. Here are some common secondary data sources used in statistics:

1. **Government agencies**: Government agencies at the local, national, and international levels collect and maintain a vast amount of statistical data. Examples include census data, labor statistics, economic indicators, crime rates, health statistics, and demographic information. These datasets are often publicly available and can provide valuable insights into various social, economic, and demographic trends.

2. **Research organizations and institutes**: Many research organizations and institutes conduct surveys, studies, and data collection efforts for specific research purposes. These organizations may focus on topics such as education, public health, social issues, or specific industries. Their datasets can provide detailed information on specific domains or research areas.

3. **International organizations**: International organizations, such as the World Bank,

International Monetary Fund (IMF), United Nations (UN), and World Health Organization (WHO), collect and maintain extensive datasets on global development, economics, health, and social indicators. These datasets cover a wide range of countries and can be used for comparative analysis and cross-country studies.

4. **Academic institutions**: Universities and research institutions often conduct research studies and surveys, resulting in datasets that can be valuable for statistical analysis. These datasets may cover various disciplines, including social sciences, psychology, economics, education, and more. Academic institutions often make their datasets available to researchers, subject to certain restrictions and ethical considerations.
5. **Nonprofit organizations**: Nonprofit organizations focused on specific causes or social issues often collect data related to their mission. These organizations may conduct surveys, compile reports, or collaborate with other entities to collect data. Their datasets can provide insights into areas such as poverty, environmental issues, human rights, and social justice.
6. **Commercial data providers**: There are commercial entities that collect, aggregate, and sell datasets on various industries, market trends, consumer behavior, and more. These datasets can be useful for market research, business analytics, and understanding consumer preferences and trends.
7. **Online platforms and social media:** Online platforms and social media networks generate vast amounts of data. This data includes user-generated content, interactions, behaviors, and demographic information. While accessing and analyzing this data may require specific permissions and compliance with privacy regulations, it can offer insights into online behavior, sentiment analysis, and social network analysis.

When using secondary data sources, researchers should consider factors such as the data quality, reliability, representativeness, and potential limitations or biases. It is essential to critically evaluate the data source and ensure that it aligns with the research objectives and analytical requirements

1.12 LIMITATIONS OF SECONDARY DATA

Although the secondary source is cheap in terms of money, time and effort, utmost care should be taken in their use. It is desirable that such data should be vast and reliable and the terms and definitions must match the terms and definitions of the current inquiry. The suitability of the data may be judged by comparing the nature and scope of the present inquiry with that of the original inquiry. Secondary data will be reliable if these were collected by unbiased, intelligent and trained investigators. The time period to which these data belong should also be properly scrutinized.

Secondary data refers to data that is collected by someone else for a different purpose but can be utilized for research or analysis. While secondary data can be convenient and cost-effective, it also has certain limitations. Here are some common limitations of secondary data collection:

1. **Lack of control over data collection:** Since secondary data is collected by others, researchers have no control over the data collection process. This can result in data that may not perfectly align with the research objectives or may lack specific variables or measures that the researcher requires. The data may not have been collected with the same level of rigor or precision as desired.

2. **Data relevance and accuracy**: The relevance and accuracy of secondary data can vary. It may be challenging to find secondary data that precisely matches the research needs, as the data may be outdated or collected using different methodologies. In some cases, the data may contain errors, inconsistencies, or missing values, which can affect its reliability and validity.

3. **Limited contextual information**: Secondary data may lack detailed information about the context in which it was collected. Understanding the specific circumstances, conditions, or nuances surrounding the data collection process may be crucial for accurate interpretation and analysis. Without sufficient contextual information, the researcher may face challenges in fully understanding and interpreting the data.

4. **Potential bias and validity concerns**: Secondary data may contain inherent biases or limitations introduced by the original data collection process. The biases could be due to the research design, sampling methods, or data collection instruments used. Researchers must critically evaluate the reliability and validity of the secondary data source to ensure its suitability for their research objectives.

5. **Incompatibility and inconsistency**: When working with secondary data from multiple sources, researchers may encounter issues of incompatibility and inconsistency. The data may have been collected using different formats, classifications, or units of measurement, making

it challenging to combine or compare the data effectively. Harmonization or standardization efforts may be necessary to address these issues.

6. **Limited control over variables**: Secondary data may not include all the variables of interest to the researcher. Certain variables that are critical for the research objectives may be missing, limiting the scope of analysis or preventing the investigation of specific relationships or factors.
7. **Data availability and access**: Accessing certain types of secondary data can be challenging due to restrictions, copyright issues, or proprietary considerations. Researchers may face limitations in obtaining the specific data they need or may need to rely on aggregated or summarized data, which may not provide the level of detail required for the research.

Despite these limitations, secondary data can still be a valuable resource for researchers, providing a foundation for analysis, hypothesis generation, and comparison with primary data. Researchers should critically evaluate the quality and relevance of the secondary data and consider its limitations in the interpretation and analysis process.

1.13 PRECAUTIONS TO COLLECT SECONDARY DATA

According to Prof. A.L. Bowley, "It is never safe to take the published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticize the arguments that can be based upon them." In using secondary data, we should take a special note of the following factors.

1) Reliable, 2) Suitable, and 3) Adequate.

Firstly, reliability of data has to be the obvious requirement of any data, and more so of secondary data. The user must make himself/herself sure about it. For this (s)he must check whether data were collected by reliable, trained and unbiased investigators from dependable sources or not.

Second, we should see whether data belong to almost the same type of class of people or not. (1) To look at and compare the given inquiry's objectives, nature, and scope with the original research. To verify that all of the terms and units were uniformly defined throughout the previous investigation and that these definitions are appropriate for the current investigation as well. For instance, a unit can be defined in multiple ways depending on its context, such as a household, wage, price, farm, etc. The secondary data will be considered inappropriate for the present research if the units were identified differently in the original investigation than what we want. lastly, consider the variations in data collecting periods and consistency of conditions comparing the original investigation and the present investigation.

Third, even if the secondary data are reliable and suitable in, it might not be adequate for the particular inquiry's objectives. This happens if the original data refers to an area or a period which is much larger or smaller than the needed one, or when the coverage given in the initial research was too narrow or too wide than what is desired in the current research. Therefore, it is make sure that due to the gap of time, the conditions prevailing then are not much different from the conditions of today in respect of habits, customs, fashion, etc. Of course, we cannot hope to get exactly the same conditions.

Suitability of data is another requirement. The research worker must ensure that the secondary data he plans to use suits his inquiry. He must match class of people, geographical area, definitions of concepts, unit of measurement, time and other such parameters of the source he wants to use with those of his inquiry. Not only this, the aim and objectives should also be matched for suitability.

Secondary data should not only be reliable and suitable, but also adequate for the present inquiry. It is always desirable that the available data be much more than required by the inquiry. For example, data on, say, consumption pattern of a state cannot be derived from the data on its major cities and towns.

CHECK YOUR PROGRESS (C)

Q1. Explain the Method to collect secondary data.

Ans.____

____

Q2. Define the Mailed questionnaire method and schedules sent through enumerator Give two limitations of primary data

Ans.____

____

Q3. Give two limitations of secondary data

Ans.____

____

1.14 SUM UP

Data Collection Methods Data / Statistics are quantitative information and can be distinguished as sample or census data; primary or secondary data. We require information for an investigation that can be gathered from either a primary source or a secondary source. Both require statistical surveys, which have two stages: planning and execution. The investigator should choose the primary or secondary sources, census or sample inquiry, type of statistical units and measurement units, level of precision desired, and other factors during the design stage. In the execution stage, the chief investigator has to set up administration, select and train field staff and supervise the entire process of data collection. Using secondary data from published or unpublished sources requires caution because they can lead to a number of problems. The questionnaire method is the most crucial of all survey methods.

A questionnaire provides a list of relevant inquiries, which should be short, clear, and of the Yes/No variety with illustrative responses. They shouldn't have a lengthy list. Questions that are private or humiliating should be avoided.

1.15 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define Business Research

Q2. Research plan

Q3. Define Primary data

Q4. Define Secondary data

Q5. What is sample

Q6. Define quantitative data

Q7. Explain qualitative data

Q8. Primary data

Q9. Secondary data

B. Long Answer Type Questions

Q1.What are the techniques for the collection of data

Q2.What are the sources of primary data?

Q3.What is a questioner? What are the points to keep in mind before drafting questioner?

Q4.Explain the term secondary data with its sources

Q5.limitations of primary data and secondary data

Q6.What are the precautions to collecting secondary Data?

1.16 SUGGESTED READINGS

- A. Abebe, J. Daniels, J.W. Mckean, “Statistics and Data Analysis”.
- Clarke, G.M. & Cooke, D., “A Basic course in Statistics”, Arnold.
- David M. Lane, “Introduction to Statistics”.
- S.C. Gupta and V.K. Kapoor, “Fundamentals of Mathematical Statistics”, Sultan Chand & Sons, New Delhi.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 2: SAMPLE MEASUREMENT AND SCALING TECHNIQUES

STRUCTURE

2.0 LEARNING OBJECTIVES

After reading the unit, learners will be able to understand:

- the concept of scales of measurement
- various properties of scales
- four types of measurement scales
- distinguish between comparative and non-comparative measurement scales and their different subtypes
- types of validity
- to get acquainted with the classification of reliability
- awareness of the measurement errors at the four hierarchical levels of a study

2.1 INTRODUCTION

Individuals and corporate entities each have unique traits that differ from one another. Humans have both concrete or quantitative attributes, such as height, weight, and complexion, as well as more abstract or qualitative traits, such as morality, creativity, attitude, and intelligence. A business organization contains tangible traits like employees, sales, offices, etc. much like human beings do. These can be easily measured because they are physical in nature.

However, there are some traits (sometimes referred to as constructs), such as credibility, an organization's perception, drive, commitment, workplace culture, and trust. All these opinions and sentiments held by clients and staff members are crucial for the business's survival and expansion. The following concepts relating to employees and customers must therefore be considered by the businesses.

Raw data is the information gathered via censuses, surveys, or other sources. Data simply means information. Before anything makes sense to be used profitably, it must be transformed into a more acceptable form. Before any conclusions can be derived from raw data, it must be transformed into the right form, such as tabulation, frequency distribution form, etc. There are two methods for gathering statistical information: Primary data and Secondary data. Primary data is information that has been gathered from a primary source by an individual or group for their benefit. Secondary data is information that has been gathered by another individual or group for their purposes, but the investigator also obtains it for his purposes.

2.2 MEASUREMENT IN RESEARCH

Associating numbers or symbols with observations found during a research investigation is known as measurement. Scales like those for hours, meters, grams, etc. can be used for this. It is quite challenging to assess or quantify motivation, for instance, if one needs to. This can be accomplished by putting motivation on a scale and assigning it some numbers. The act of measuring involves keeping track of observations made as part of a research project. By certain guidelines, the observations may be recorded using numbers or other symbols to represent an object's qualities. Characteristics of the respondents include their thoughts, feelings, and behaviors. Assigning '1' for male responses and '2' for female respondents is one example.

The respondent has the option of answering "yes" or "no" when asked whether they use the "Internet Banking" service offered by a specific bank branch. You could want to use the numbers "1" for "yes" and "2" for "no" as your response codes. For two reasons, we assign numbers to these traits. First, the numbers make it easier to conduct additional statistical analysis of the data. Second, statistics make it easier to communicate measuring guidelines and outcomes. The definition of guidelines for allocating numbers to attributes is the most crucial component of measurement. Numbering conventions ought to be standardized and implemented consistently. This cannot alter as time or things change.

2.3 SCALING IN RESEARCH

The process of converting a group of textual statements to numbers by a rule is known as scaling. The objects in scaling are textual assertions, typically ones that express attitude, opinion, or emotion. Consider, for instance, a scale that categorizes bank customers based on whether they "agree to the satisfactory quality of service provided by the branch." Each consumer who is surveyed has the option of responding with one of the following phrases: "strongly agree," "somewhat agree," "somewhat disagree," or "strongly disagree." We might even give a number to each response. For instance, we might rate each response as strongly agreeing at 1, agreeing at 2, disagreeing at 3, and severely disagreeing at 4. As a result, each respondent may choose from 1, 2, 3, or 4.

2.4 PROPERTIES OF SCALES

a) Distinctive Classification: A measure is considered to have this attribute if it can be used to divide objects or their properties into discrete classes or categories. This is a prerequisite for every metric. For instance, gender divides the population into two separate categories males and females.

b) Order: If the components of a measure can be placed in a logical order, then the measure is said to have ordered. For instance, a student's grades may be arranged in either an ascending or downward order.

c) Equal Distance: A measure is said to have an equal distance if there is an equal distance between any two consecutive categories of a measured property (usually referred to as values for numeric variables). For instance, the time difference between 2 and 3 in the afternoon is the same as the time difference between 3 and 4 in the afternoon, or 1 hour.

d) **Fixed Origin:** If there is a meaningful zero or an 'absence' of the characteristic, the measurement scale for that characteristic is said to have a fixed origin. Examples include a person's income, a business's sales, etc.

2.5 NOMINAL SCALE

Nominal scales are qualitative scales without any sort of order. This scale only fits into one category and does not meet the other three criteria listed above. It is referred to as "nominal" because, even though one may use numbers to represent the categories, these numbers are merely "nominal" and have no intrinsic value, order, or significance. An example of a nominal measure is the color of bicycles. Which hue would you want for a bike? has a variety of potential options, including blue, black, red, etc. These hues can be numbered 1, 2, 3, or 4 in any order; this scale neither has a set order nor a value. These categorized terms are given numerical values. These codes are employed to identify people. Nominal data refers to information gathered using a scale of nominal measures. Nominal scale data is of a form that may be categorized into groups or categories and given names to describe them. Examples include a person's address, phone number, vehicle identification number, and entrance exam applicant roll number. Sometimes, instead of using numbers to categorize things, codes are used, such as STD codes for cities, bar codes for things in department shops, codes for different university disciplines, codes for books in libraries, blood type codes for people, etc.

2.6 ORDINAL SCALE

The identity and magnitude properties of the ordinal scale exist. Each value on the ordinal scale has a distinct meaning and is related to the other values in an ordered manner. Ordinal scales are used to assess things like attitudes and preferences as well as occupations and social classes. Ordinal scales assist in putting various elements, such as items, people, or reactions, in relative positions about a specific aspect. It is a rating scale in which items are given numbers to show the relative degree to which they contain a particular attribute. It can determine whether an object has a characteristic more or less than another object, but not by how much. For instance, if you were to imagine a competition, you could rank the winners from first to last. If someone says they placed second, you know that one competitor placed first and everyone else placed second. Ordinal variables, however, don't provide any information regarding the exact size of the gap between first and second or between second and third. Let's look at a few instances: Rank the following characteristics in order of significance (1–5) when buying a desktop.

1. Brand Name
2. Functions
3. Price
4. After-sale services
5. Design

The respondents ranked each attribute from 1 to 5, with 5 being the least important. To rate on an ordinal scale, letters or symbols may also be used in place of numbers. Such a scale does not attempt to quantify the degree to which certain rankings are favorable. The use of an ordinal scale is once more necessary if there are four different types of pesticides and if they are ranked according to quality as Grade A, Grade B, Grade C, and Grade D. Ordinal scales enable the application of statistics like percentile, quartile, and more in addition to the counting operation allowed for data on a nominal scale. A mean cannot be utilized only a mode or median may.

2.7 INTERVAL SCALE

An interval scale is a measurement system whose base value is not set and whose succeeding values indicate equal amounts or values of the feature being measured. There is no genuine or fixed zero on this quantitative scale of measurement. Interval data refers to information gathered using an interval scale. Quantitative information that may be assessed on a scale of 1 to 10 is interval data. The zero point, however, does not imply that the trait being assessed is not present. Temperature, time, longitude, latitude, etc. are a few examples. The Fahrenheit temperature scale is an illustration of an interval scale. The temperature difference between 20 degrees and 40 degrees Fahrenheit is equivalent to the difference between 75 degrees and 95 degrees. There is no absolute zero with interval scales. It would not be permissible to suggest that 60 degrees is twice as hot as 30 degrees because it would be improper to express Interval level measurements as ratios. To get the average scores for each attribute across all respondents, the data from the interval scale can be used. You can also compute the Standard Deviation, which is a measure of dispersion. Common statistical measurements like range, standard deviation, and correlation are all measured using this scale. The distance between each point on the scale is the same as the researcher continuously measures the preference, liking, or importance of a specific brand attribute. The zero point's placement is movable. The units of measurement and the zero point are both arbitrary. On interval scaled data, you can perform bivariate correlation analyses, t-tests, analysis of variance tests, and most multivariate techniques used for inference drawing.

2.8 RATIO SCALE

The greatest level of measurement scale is the ratio scale. This has a fixed (absolute) zero point and the characteristics of an interval scale. We can create a meaningful ratio, thanks to the absolute zero point. Scales that use ratios include those that use weights, lengths, and times. An example of a ratio scale is the volume of ATM users over the last three months for a bank. This is so that you may evaluate it in comparison to the prior three months. The researcher can compare both disparities in scores and the relative magnitude of scores using ratio scales. For instance, the time difference between 10 and 15 minutes and the time difference between 25 and 30 minutes are equal, however, the time difference between 15 and 30 minutes is twice as long. Ratio scales are commonly used in financial research that examines rupee values. However, interval scales are often the best type of assessment for most behavioral studies.

2.9 COMPARATIVE SCALES

The various objects are directly compared using the comparative scales. For instance, in research of customer preferences for several airlines, a consumer may be asked to rate a list of variables, such as price, punctuality, food, a flying returns program, etc., that he or she would consider when choosing a specific airline. The most favoured factor must be ranked as number one, and the least preferred factor must be ranked last.

2.9.1 Paired Comparison Scale: A respondent is given two objects at once and asked to choose one (rate between two objects at once) based on a set of criteria in this comparative scaling technique. The collected data are ordinal. We typically have n (n - 1)/2 paired comparisons for n brands. The data recording format for paired comparisons is as follows.

The following paired comparisons on three parameters were requested as part of a research of customer preferences on two brands of Chocolate, KitKat and Five Star. Simply pick one brand out of the two.

- Which Chocolate do you prefer based on 'TASTE'?

  KitKat Five Star
- Which Glucose biscuits do you prefer based on 'PRICE'?

  Sunfeast Parle G
- Which Glucose biscuits do you prefer based on 'PACKAGING'?

  Parle G Tiger Biscuits

2.9.2 Rank Order Scale: This is a different kind of comparative scaling technique where respondents are shown multiple items at once and asked to rank them in terms of importance. This ordinal scale describes the preferred and disfavoured objects but conceals the separation between them. For instance, you might use the following format to record the replies if you were interested in ranking the preferences of a few chosen brands of cold drinks. The rank order scale is likewise comparable in nature, much like paired comparison. Ordinal data are the results in rank order. When direct comparisons between the provided objects are necessary, this technique produces superior results since it is more realistic in producing the responses. The main drawback of this method is that it can only produce ordinal data.

Example: When choosing a new mobile service provider, rank the following services in the order of significance that you assign to them. Rankings can start at 1, move up to 2, and so forth.

Feature	Rank
1. Connectivity	-----
2. Minimum Call Drops	-----
3. Internet	-----
4. Value Added Services	-----
5. Roaming	-----
6. Ring tone/Caller tune	-----
7. Alerts	-----
8. Downloads	----
9. SMS	----

2.9.3 Constant Sum Scale: According to a criterion, respondents are asked to distribute a fixed number of units, such as points, rupees, or chips, among a group of stimuli. For instance, you might want to research how essential consumers think a detergent's price, aroma, packaging, cleaning ability, and lather are. The following structure may be used to ask respondents to indicate the relative importance of the traits by dividing a fixed total. Cleaning ability and packaging rank second and third in importance for consumers. The two qualities that people are concerned about the least yet prefer equally are fragrance and lather. Saving time is a benefit of this method. There are two significant drawbacks, though. The use of too few attributes might result in rounding off mistakes, while using too many attributes may be excessively stressful on the respondent and lead to confusion and tiredness. Decide how much of the total of Rs. 6000 you would want to spend on the following things on your birthday (please note that the total money allocated must equal exactly 6000).

Item	Amount
1. Cosmetics	-----
2. Clothes	-----
3 Accessories	-----
4. Jewelry	-----
5. Dinner	-----
6. Movie	-----
Total	6000/-

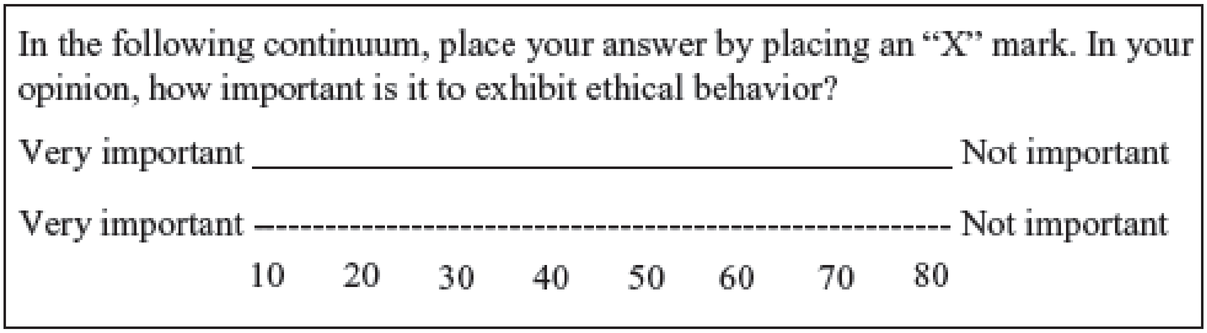
Item Amount 1. Cosmetics _____ 2. _____ 3. _____ 4. _____ 5. _____ 6. _____ ______________________________________ Total 5000

2.9.4 Q-Sort Scale: This comparison scale sorts items depending on how similar they are to some criterion using a rank order approach. The key feature of this research is that comparing responses from distinct respondents is less significant than comparing responses from different respondents' responses. As a result, rather than being an absolute rating scale, it uses a comparative approach of scaling. In this procedure, the respondent is given a huge number of statements that describe a product's features or those of numerous different brands of the same product.

2.10 NON-COMPARATIVE SCALES

Respondents in noncomparative scaling only need to evaluate one object. Their assessment is separate from the researcher's analysis of the other object. When utilizing a non-comparative scale, respondents use any rating criterion they see acceptable. Continuous and itemized rating scales are non-comparative methods.

2.10.1 Continuous Rating Scales: It is straightforward and very helpful. According to a continuous line that extends from one extreme of the criterion variable to the other, the respondent rates the objects by placing a mark in the proper location on the line.



In the following continuum, place your answer by placing an "X" mark. In your opinion, how important is it to exhibit ethical behavior?

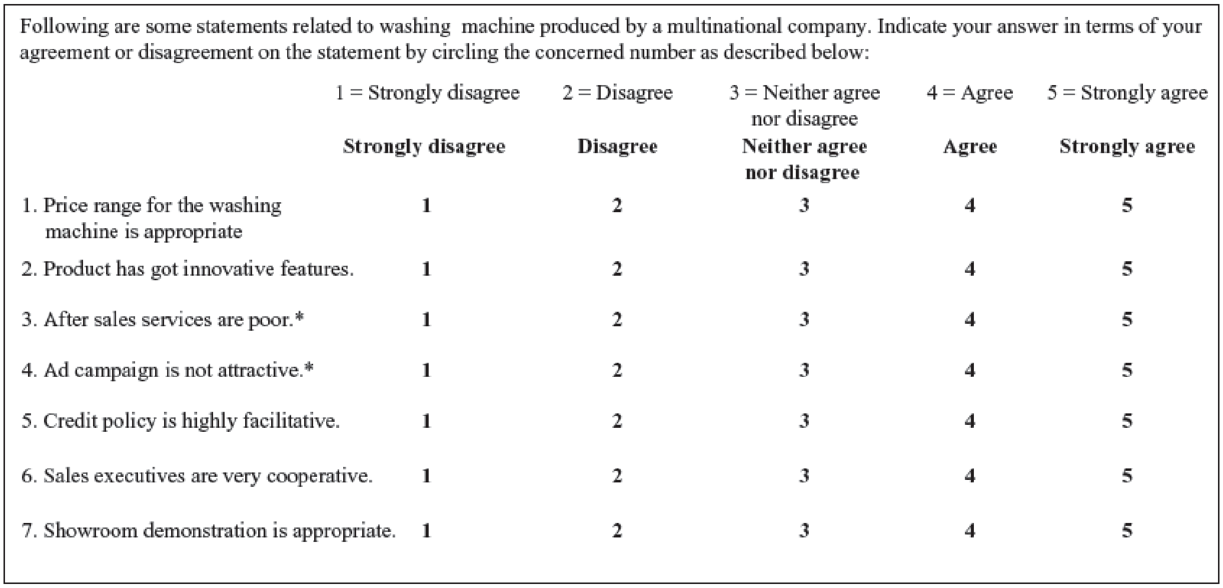
Very important ____ Not important

Very important --- Not important

10 20 30 40 50 60 70 80

2.10.2 Itemized Rating Scales: The scale with numbers or brief descriptions assigned to each category is known as an itemized rating scale. The respondents are asked to choose one of the few categories that best describes the product, brand, company, or product feature being scored. The categories are arranged according to scale position. In marketing research, itemized rating scales are frequently employed. In this section below we will discuss three itemized rating scales, namely (a) Likert scale, (b) Semantic Differential Scale, and (c) Stapel Scale.

(a) Likert Scale: Because it is so easy to use, Rensis Likert's scale has become very common in business research for evaluating attitudes. By marking how strongly they agree or disagree with carefully crafted phrases that vary from highly positive to very negative towards the attitudinal object, the respondents use the Likert scale to express their attitudes. Typically, respondents have five options to pick from: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree. Other variations in Likert scale are that these can be 7-point and 9-point scales too.



Following are some statements related to washing machine produced by a multinational company. Indicate your answer in terms of your agreement or disagreement on the statement by circling the concerned number as described below:

	1 = Strongly disagree	2 = Disagree	3 = Neither agree nor disagree	4 = Agree	5 = Strongly agree
	Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
1. Price range for the washing machine is appropriate	1	2	3	4	5
2. Product has got innovative features.	1	2	3	4	5
3. After sales services are poor.*	1	2	3	4	5
4. Ad campaign is not attractive.*	1	2	3	4	5
5. Credit policy is highly facilitative.	1	2	3	4	5
6. Sales executives are very cooperative.	1	2	3	4	5
7. Showroom demonstration is appropriate.	1	2	3	4	5

(b) Semantic Differential Scale: This rating scale has seven points, and the endpoints are bipolar labels with semantic significance (such as good and poor, complex and simple, Optimistic and Pessimistic, Introvert and Extrovert etc.). There are several uses for the Semantic Differential scale. It can be used to determine whether a respondent has a favorable or unfavorable opinion of an object. It has been extensively utilized when contrasting brands, goods, and company reputations. Additionally, it has been applied in a study on new product development as well as the creation of advertising and marketing tactics.

Rate the ATM you have just used in respect of the indicated parameters. Mark × at an appropriate location that best suits your answer.

- The ATM was _____________ for operations.
- Easy: ___: ___:___:___:___:___:___: Difficult
- The processing time was Slow: ___: ___:___:___:___:___:___: Fast
- The security person was Cordial: ___: ___:___:___:___:___:___: Indifferent

(c) Staple Scale: Staple scale is an 11-point scale where +5 and -5 are assigned above and below a factor or feature of a product etc.

Rate the outlet on the following factors. +5 indicates that the factor is most accurate for you and – 5 indicates that the factor is most inaccurate for you.		
+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
Good Ambience	Quality Products	Excellent Service
-1	-1	-1
-2	-2	-2
-3	-3	-3
-4	-4	-4
-5	-5	-5

2.11 VALIDITY

An instrument's validity is its capacity to measure what it is intended to measure. The idea that a measure should measure what it is designed to measure is straightforward in theory but extremely challenging in practice. If a scale is to measure Emotional Quotient (E.Q.) then all the items should measure EQ only not any other construct.

a) **Content Validity:** The meticulous specification of constructs, the examination of scaling techniques by content validity judges, and engagement with experts and members of the population are only a few examples of the content validation (Vogt et al., 2004). The term "face validity" may also be used to refer to content validity. In actuality, the scale's content validity is a subjective assessment of its capacity to measure what it is intended to measure.

b) **Construct Validity:** The first idea, question, or hypothesis that establishes which data should be created and how they should be collected is known as construct validity (Golafshani, 2003). The researcher must concentrate on convergent validity and discriminant validity to obtain construct validity. When the new measure correlates or converges with other comparable measures, the convergent validity is established. The definition of correlation or convergence in its literal sense refers to how closely a score on one scale correlates with a score on another scale created to measure the same constructs.

c) **Discriminant validity:** When a new measuring tool exhibits low correlation or non-convergence with the measurements of distinct concepts, discriminant validity is proven. The degree to which the score on one measuring instrument (or scale) is not correlated with the other measuring instrument (or scale) established to assess the distinct constructs is the literal meaning of the terms "no correlation" or "non-convergence." A researcher must establish the convergent validity and discriminant validity before they can establish the construct validity.

2.12 RELIABILITY

According to Burns and Bush (2016), reliability is the propensity of a respondent to give the same or a comparable response to a question that is almost identical to another inquiry. When the same person responds the same way when the same measuring device is provided to them repeatedly under the same or nearly identical conditions, that measurement is dependable. Reliable measuring tools assure researchers that temporary and contextual elements are not interfering with the process, and the measuring tool is therefore resilient. Test-retest reliability, equivalent form reliability, and internal consistency reliability are the three strategies a researcher might use to address the dependability issue.

a) **Test-Retest Reliability:** The same respondents are given the identical questionnaire to elicit replies in two distinct time slots to carry out the test-retest reliability. The degree of similarity between the two groups of responses is then calculated as a further step. The correlation coefficient is calculated to determine how similar the two sets of responses are to one another. A more dependable measuring instrument is one with a greater correlation coefficient, while the opposite is true for unreliable instruments.

b) **Equivalent Forms Reliability:** While considering the dependability of similar forms, two identical forms are given to the participants at two separate times, and test-retest reliability considers personal and situational variation in responses over two different periods. Two similar forms are created using various item samples to measure the desired features of interest. Both formats have the same structure and type of questions, with a few key variations.

c) **Internal Consistency Reliability:** A summated scale, in which many items are added together to generate a single score, is evaluated for reliability using internal consistency reliability

(Malhotra, 2004). Split-half technique is the fundamental method to assess the internal consistency reliability. With this method, the objects are separated into equal groups. This divide is made based on some established factors, such as the questionnaire's odd-versus-even questions, or the splitting of the items at random. Responses to items are associated after division. High internal consistency is indicated by a high correlation coefficient, and low internal consistency is shown by a low correlation coefficient. Researchers frequently run into issues when dividing materials into two portions because of subjectivity. Coefficient alpha, often known as Cronbach's alpha, is a strategy that is frequently used to address this issue.

2.13 MEASUREMENT ERRORS

a) The measurement inaccuracy that poses the greatest challenge includes a stable characteristic of the object or event in addition to the one that the researcher is interested in. The measuring of attitudinal reactions has been proven to be biased by extraneous factors like gender, education, age, etc. When they vary amongst respondent groups of interest, stable features of respondents—such as gender, culture, subculture, nationality, etc.—can be particularly problematic. For instance, it is rare to hear a direct "no" from a Japanese person.

b) The influence of an object's short-term properties is another prevalent source of inaccuracy. Such factors as fatigue, health, hunger, and emotional state may influence the measure of the other characteristics. For instance, a person's responses may change if they are feeling downhearted due to a cold or exhaustion.

c) Numerous measurements that include human beings consider both the genuine characteristic being studied and the conditions under which the measurement is being made. For instance, husbands and wives frequently report having one amount of influence over a choice to buy something when their spouses are there, and another level of impact when their spouses are not.

d) The method used to collect the data may also have an impact on the measurement. An interviewee's response patterns to questions are influenced by the interviewer's gender, age, ethnicity, and clothing choice. Additionally, different forms of communication, such as telephone, mail, in-person interviews, and the like, might occasionally change response patterns. Aspects of the measuring instrument itself can cause constant or random errors. Unclear instructions, ambiguous questions, confusing terms, irrelevant questions, and omitted questions can all introduce errors.

f) Another reason why answers could not correctly reflect the 'actual' characteristic is response errors. A respondent might have unintentionally checked a good response when they meant to check a negative one, for instance.

g) Last but not least, errors might be made when deciphering, coding, tabulating, and analyzing a person's or a group's response. As an illustration, the researcher might enter 8 as a response rather than 3.

2.14 SUM UP

Measurement is the process of relating numerical or graphical representations to the observations made during a research project. Scaling is the process of assigning items to numbers or meanings in line with a rule. Nominal, ordinal, interval, and ratio measurements are the four different types of levels. These scales make up a hierarchy, with the nominal scale of measurement having significantly fewer statistical applications than scales higher up the hierarchy. Scales possess the following four characteristics: distinct classification, order, equal distance, and fixed origin. Data on categories are provided by nominal scales, sequences are provided by ordinal scales, magnitudes between points on the scale are revealed by interval scales, and order and absolute distance between any two points on the scale are both explained by ratio scales. Two categories of measurement scales are frequently used in marketing research: comparative scales and non-comparative scales. To communicate differences between two or more businesses, brands, services, or other stimuli, respondents use comparative scales. The scales under this type are: (a) Paired Comparison, (b) Rank Order, (c) Constant Sum, and (d) Q-sort. Further, the non-comparative scales can be classified into: (a) Continuous Rating Scales and (b) Itemized Rating Scales. The Itemized Rating scales can further be classified into: (a) the Likert Scale, (b) Semantic Differential Scale, and (c) the Stapel Scale. There are numerous scaling methods available for measuring attitudes. Reliability is the likelihood of a respondent to give the same or a comparable response to a question that is practically identical to another inquiry. An instrument's validity is its capacity to measure what it was designed to measure.

2.15 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define measurement in Research.

Q2. Name four fundamental scales in Research Parlance.

Q3. Define Validity.

Q4. Define Reliability.

Q5. What is Cronbach's Alpha?

Q6. What is the Staple scale?

Q7. Enumerate various Comparative Scales.

Q8. Enlist various non-comparative scales.

Q9. What is Construct validity?

Q10. Define Measurement errors.

B. Long Answer Type Questions

Q1. Explain in brief the concept of measurement in research.

Q2. What do you understand by "Scaling" in research?

Q3. Differentiate with the help of examples between nominal, ordinal, interval and ratio scale

Q4. Differentiate between ranking scales and rating scales. Which one of these scales is better for measuring attitudes?

Q5. Name any four situations in commerce where you can use the Likert scale.

Q6. Point out the possible sources of error in measurement. Describe the tests of sound measurement.

Q7. Discuss the relative merits and demerits of Summated and Cumulative scales.

Q8. "Validity is more critical to measurement than reliability". Comment.

Q9. "Reliable measurement is necessarily a valid measurement". Comment.

Q10. A researcher wants to measure consumer preference between 9 brands of vegetable oil and has decided to use the paired comparison method. How many pairs of brands will the researcher present to the respondents?

2.16 SUGGESTED READINGS

- Naresh Malhotra (2004), "Marketing Research: An Applied Orientation", Prentice Hall International Edition
- Dillon, W. R., Madden, T. S. and Firtle, N. H. (1994), Marketing Research in a Marketing Environment, 3rd edition, Irwin, p. 298.

- Aaker, David A. and George S. Day. (1983) Marketing Research, John Wiley, New York.
- Bailey, Kenneth D. (1978) Methods of Social Research, The Free Press, New York.
- Coombs, C.H. (1953) “Theory and Methods of Social Measurement”, in Research Methods in the Behavioral Sciences, eds. Feslinger, L. and Ratz, D., Holt, Rinehart and Winston.
- Donald S. Tull and Gerald S. Albaum. (1973) Survey Research: A Decisional Approach, Index Educational Publishers, New York.
- Meister, David. (1985) Behavioural Analysis and Measurement Methods, John Wiley, New York.
- Rodger, Lesile W. (1984) Statistics for Marketing, McGraw-Hill (UK), London.
- Thurstone, L. L., (1927), "A Law of Comparative Judgment", Psychological Review 34, pp. 273-86.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.
- S.P. Gupta, “Statistical methods” Sultan Chand & Sons publication, New Delhi
- Saunders, “Research Methods for Business Students”, Pearsons Education Publications.
- Burns, A., Veeck, A. and Bush, R. (2016), Marketing Research, Pearson
- Golafshani, N. (2003). Understanding Reliability and Validity in Qualitative Research, The Qualitative Report, 8(4), 597-606.
- S Vogt, D.S., King, D.W. and King, L.A. (2004), Focus groups in psychological assessment: enhancing content validity by consulting members of the target population, Psychol Assess, 16 (3), 231-43.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 3: PROCESSING AND PRESERVATION OF DATA: PROCESSING OF DATA, DIAGRAMMATIC AND GRAPHIC PRESENTATION

STRUCTURE

3.0 LEARNING OBJECTIVES

After studying the Unit, learners will be able to know:

- Processing the data after collection
- Various stages of Processing
- Plan the data analysis
- Classify the data
- Tabulate the data
- Analyse the data
- Data Preservation, storage and Retention

3.1 INTRODUCTION

After data collection, the researcher turns his focus of attention to the processing and preserving of the data. "Diagrammatic" and "Graphical" presentation of data both refer to methods of visually representing information to enhance understanding and analysis. While the terms are often used interchangeably, they can have slightly different connotations.

A diagrammatic presentation involves using diagrams or visual aids to represent data. Diagrams are usually simplified and symbolic representations that convey information clearly and straightforwardly. Common types of diagrammatic presentations include bar charts, pie charts, line graphs, histograms, scatter plots, and pictograms. These visual representations help illustrate patterns, trends, comparisons, and relationships within the data.

Graphical presentation refers to the use of graphs, charts, and visual elements to display data in a way that makes it easier to interpret and analyze. Graphs can include various types of charts, plots, and diagrams that present data in a visual format. This type of presentation is particularly useful when dealing with complex data sets or when you want to emphasize relationships and trends. Graphical presentations can include more detailed and sophisticated visualizations, such as 3D graphs, heat maps, area charts, and more.

3.2 DATA PROCESSING

Data processing refers to certain operations such as editing, coding, computing of the scores, preparation of master charts, etc. A researcher has to make a plan for every stage of the research process. As such, a good researcher makes a perfect plan for processing and analysis of data. To some researchers' data processing and analysis is not a very serious activity.

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly so as not to negatively affect the end product or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

Stages of Data Processing

1) **Data Collection:** Collecting data is the first step in data processing. Data is pulled from available sources. The data sources available must be trustworthy and well-built so the data collected (and later used as information) is of the highest possible quality.
2) **Data Preparation:** Once the data is collected, it then enters the data preparation stage. Data preparation, often referred to as "pre-processing" is the stage at which raw data is cleaned up and organized for the following stage of data processing. During preparation, raw data is diligently checked for any errors. The purpose of this step is to eliminate bad redundant, incomplete, or incorrect data and begin to create high-quality data for the best business intelligence.
3) **Data Input:** The data is then entered into its destination and translated into a language that it can understand. Data input is the first stage in which raw data begins to take the form of usable information.
4) **Processing:** During this stage, the data inputted to the computer in the previous stage is processed for interpretation. Processing is done using machine learning algorithms, though the process itself may vary slightly depending on the source of data being processed.
5) **Data output/interpretation:** The output/interpretation stage is the stage at which data is finally usable to non-data scientists. It is translated, readable, and often in the form of graphs, videos, images, plain text, etc.).
6) **Data storage:** The final stage of data processing is storage. After all of the data is processed, it is then stored for future use. While some information may be put to use immediately, much of it will serve a purpose later on. When data is properly stored, it can be quickly and easily accessed by members of the organization when needed.

3.3 DIAGRAMMATIC PRESENTATION OF DATA PROCESSING

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

➢ Rules for Preparing Diagrams

The prime objective of the diagrammatic presentation of data is to highlight their basic hidden facts and relationships. To ensure that the presentation of numerical data is more attractive and effective, therefore, it is essential to keep the following general rules in mind while adapting diagrams in research work. Now, let us discuss them one by one.

1. You must have noted that the diagrams must be geometrically accurate. Therefore, they should be drawn on the graphic axis i.e., the 'X' axis (horizontal line) and the 'Y' axis (vertical line). However, the diagrams are generally drawn on plain paper after considering the scale.
2. While taking the scale on the 'X' axis and 'Y' axis, you must ensure that the scale showing the values should be in multiples of 2, 5, 10, 20, 50, etc.
3. The scale should be set up, e.g., millions of tons, persons in Lakhs, value in thousands, etc. On the 'Y' axis the scale starts from zero, as the vertical scale is not broken.
4. Every diagram must have a concise and self-explanatory title, which may be written at the top or bottom of the diagram.
5. To draw the readers' attention, diagrams must be attractive and well-proportioned.
6. Different colours or shades should be used to exhibit various components of diagrams and also an index must be provided for identification.
7. It is essential to choose a suitable type of diagram. The selection will depend upon the number of variables, minimum and maximum values, and objects of presentation.

3.4 GRAPHICAL REPRESENTATION OF DATA PROCESSING USING EXCEL

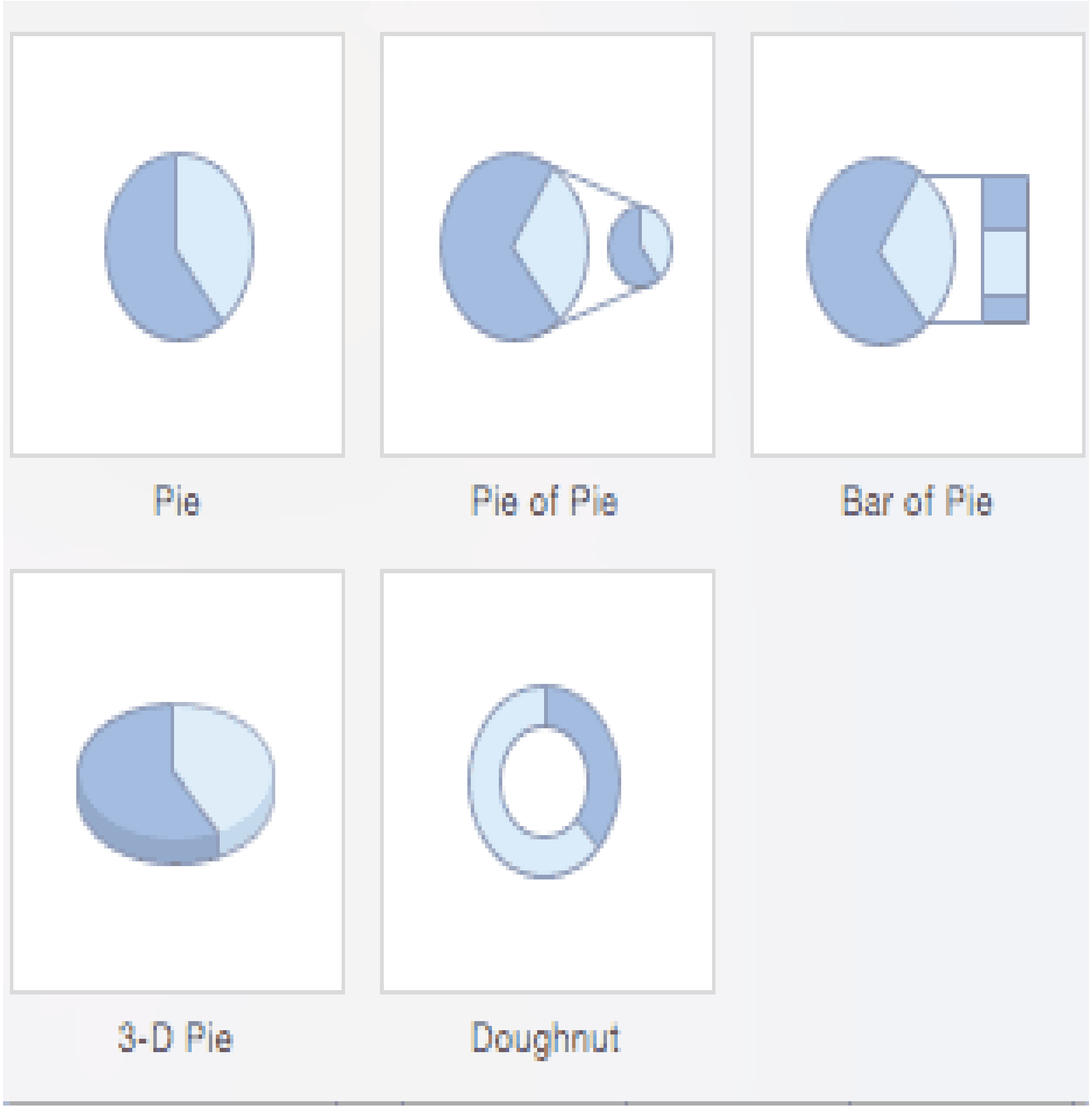
Excel charts are graphical representations of numeric data. Graphs make it easier for users to compare and understand numbers, so charts have become a popular way to present numerical data. Every chart tells a story. Stories can be simple: “See how our sales have increased” or complex: “This is how our overhead costs relate to the price of our product.” Whether simple or complex, the story should be readily understandable. If you can’t immediately understand what a chart means, then it isn’t a good chart.

Graphs are constructed with data points, which are the individual number in a worksheet, and data series, which are the groups of related data points within a column or row. Charts and graphs in Microsoft Excel provide a method to visualize numeric data. While both graphs and charts display sets of data points about one another, charts tend to be more complex, varied, and dynamic. People often use charts and graphs in presentations to give management, client, or team members a quick snapshot of progress or results. You can create a chart or graph to represent nearly any kind of quantitative data — doing so will save you the time and frustration of poring through spreadsheets to find relationships and trends. It’s easy to create charts and graphs in Excel, especially since you can also store your data directly in an Excel Workbook, rather than importing data from another program. Excel also has a variety of preset chart and graph types so you can select one that best represents the data relationship(s) you want to highlight. Excel comes with a wide variety of charts capable of graphically representing most standard types of data analysis and even some more exotic numeric interpolations. The type of data you are using and presenting determines the type of chart you will plot the data on.

3.5 TYPES OF CHARTS

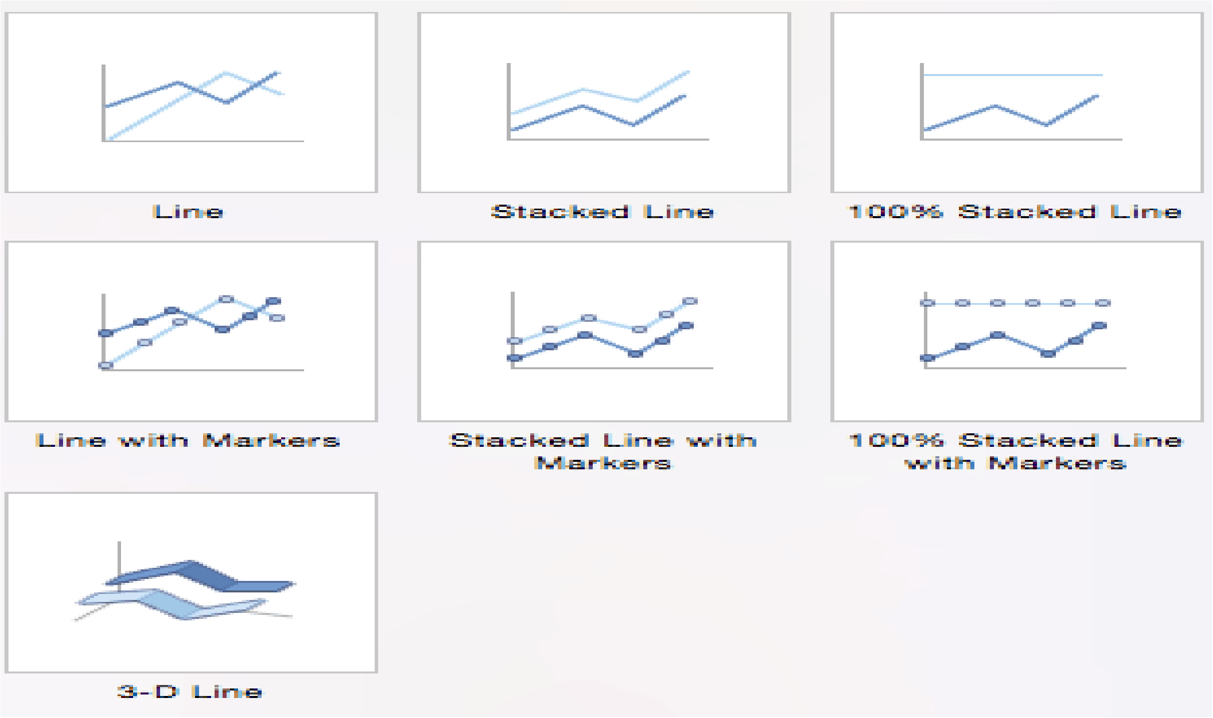
1. Pie Charts: These work best for displaying how much each part contributes to a total value. Pie charts can be exploded for greater visual clarity, or turned into doughnut charts, which can represent more than just one set of data.

- Use pie charts to compare percentages of a whole (“whole” is the total of the values in your data). Each value is represented as a piece of the pie so you can identify the proportions. There are five pie chart types: pie, pie of pie (this breaks out one piece of the pie into another pie to show its sub-category proportions), bar of pie, 3-D pie, and doughnut.

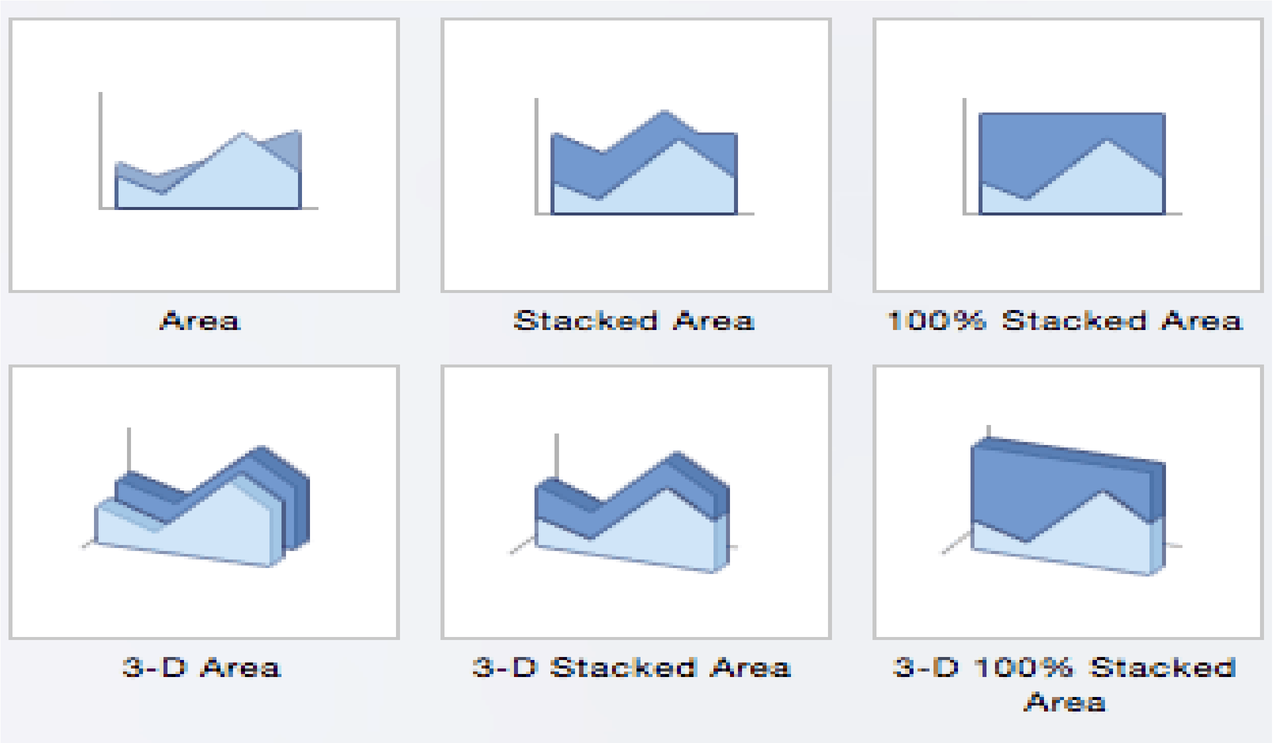


2. Line and area charts: These show data points connected with lines, indicating upward or downward trends in value. Area charts show the area below a line filled in. Both types can be combined with column charts to show more data.

- A line chart is most useful for showing trends over time, rather than static data points. The lines connect each data point so that you can see how the value(s) increased or decreased over some time. The seven-line chart options are line, stacked line, 100% stacked line, line with markers, stacked line with markers, 100% stacked line with markers, and 3-D line.

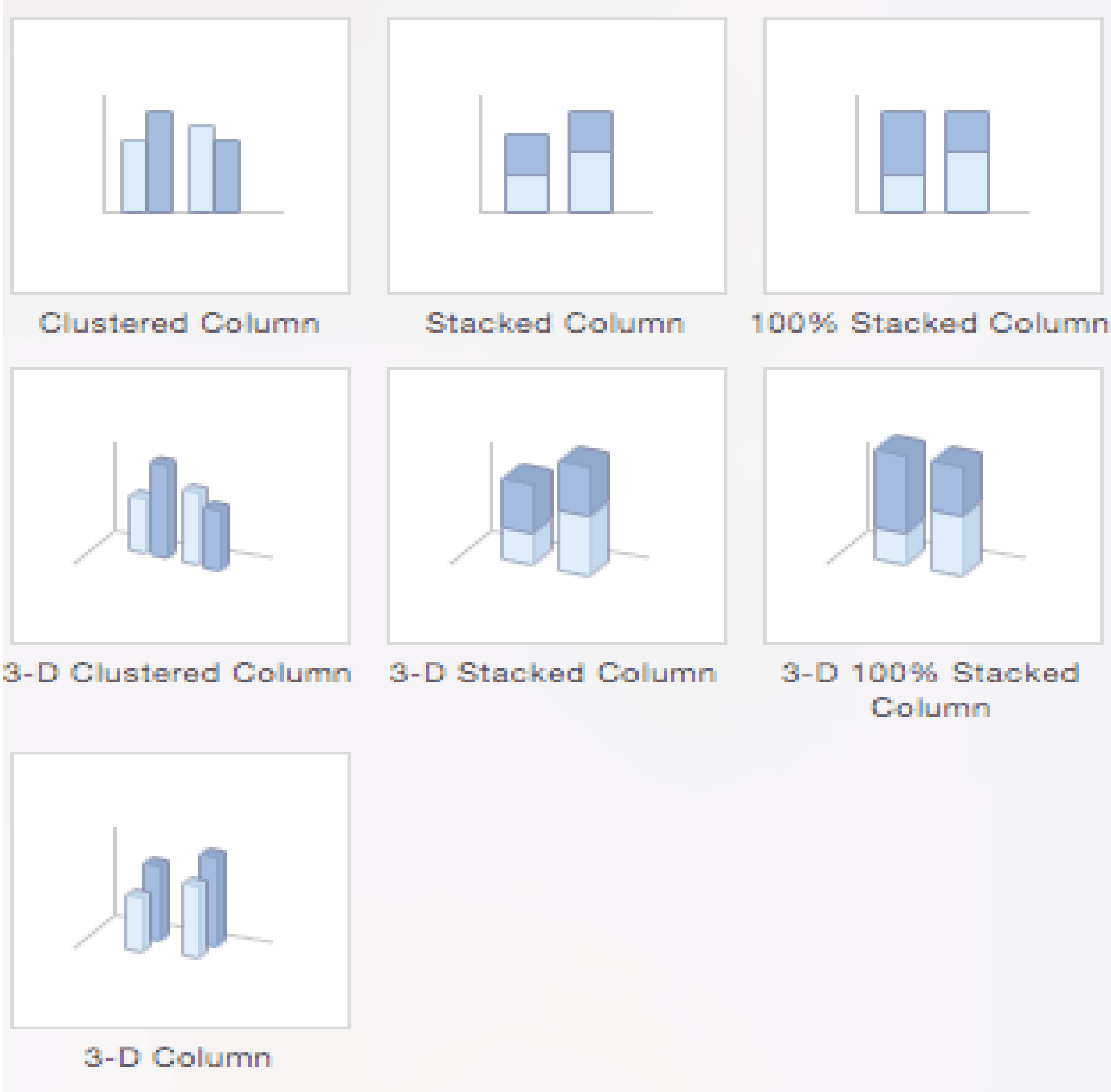


- **Area:** Like line charts, area charts show changes in values over time. However, because the area beneath each line is solid, area charts are useful to call attention to the differences in change among multiple variables. There are six area charts: area, stacked area, 100% stacked area, 3-D area, 3-D stacked area, and 3-D 100% stacked area.

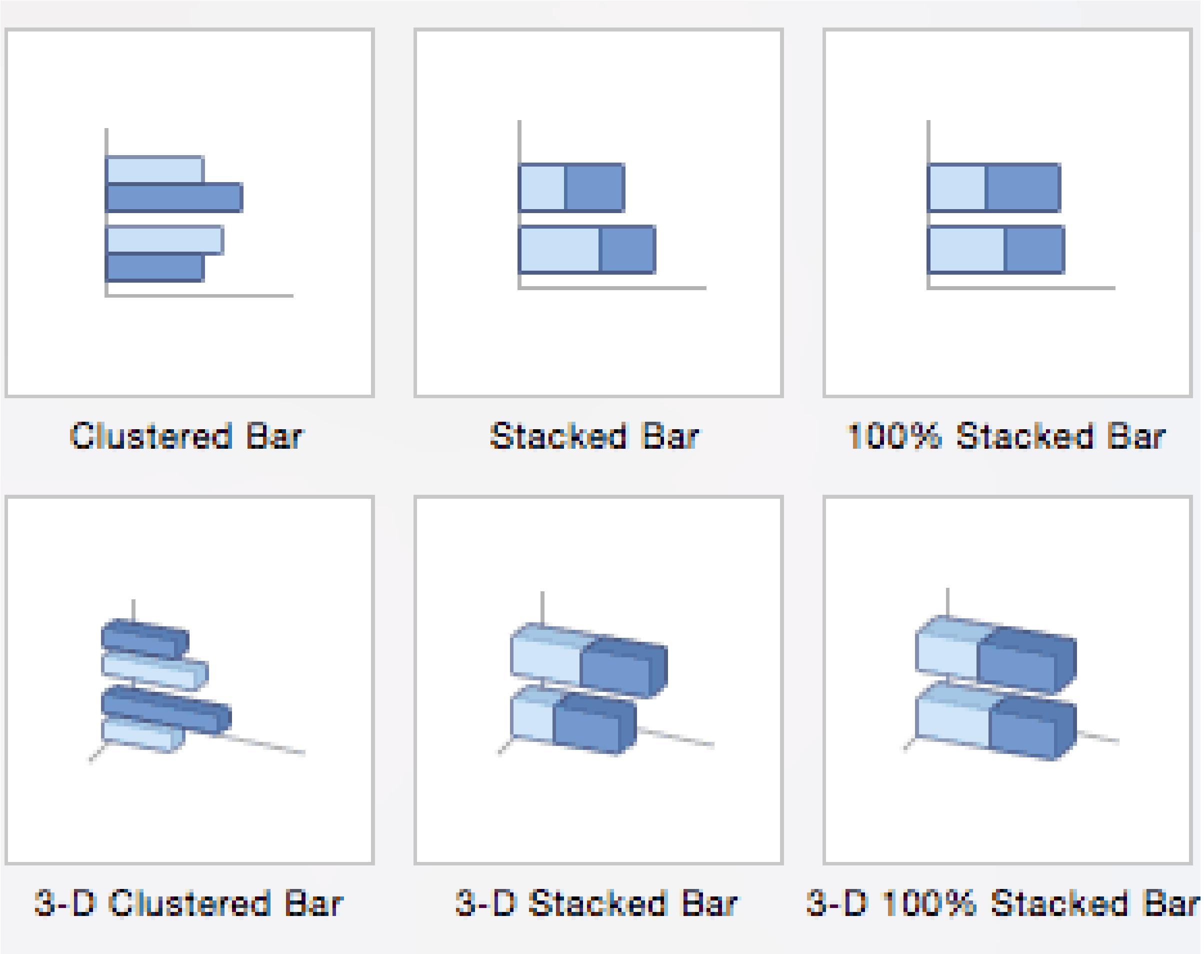


1. **Column and bar charts:** These compare values across categories, with results presented vertically in column charts and horizontally in bar charts, The composition of the column or bar can be stacked in more than one color to represent the contribution of each portion of a category's data to the total for that category.

- **Column Charts:** Some of the most commonly used charts, column charts, are best used to compare information or if you have multiple categories of one variable (for example, multiple products or genres). Excel offers seven different column chart types: clustered, stacked, 100% stacked, 3-D clustered, 3-D stacked, 3-D 100% stacked, and 3-D, pictured below. Pick the visualization that will best tell your data's story.

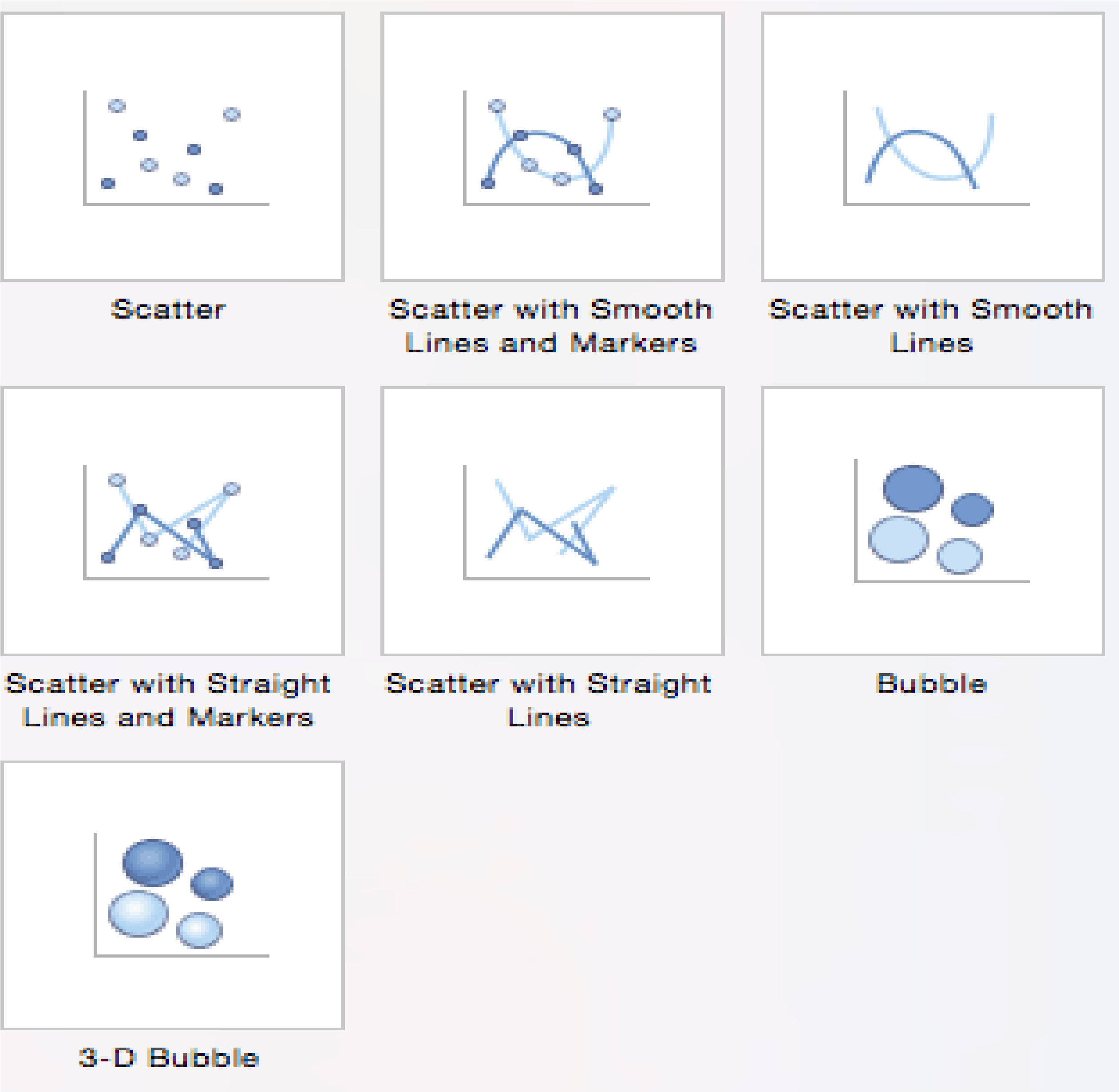


- **Bar Charts:** The main difference between bar charts and column charts is that the bars are horizontal instead of vertical. You can often use bar charts interchangeably with column charts, although some prefer column charts when working with negative values because it is easier to visualize negatives vertically, on a y-axis.

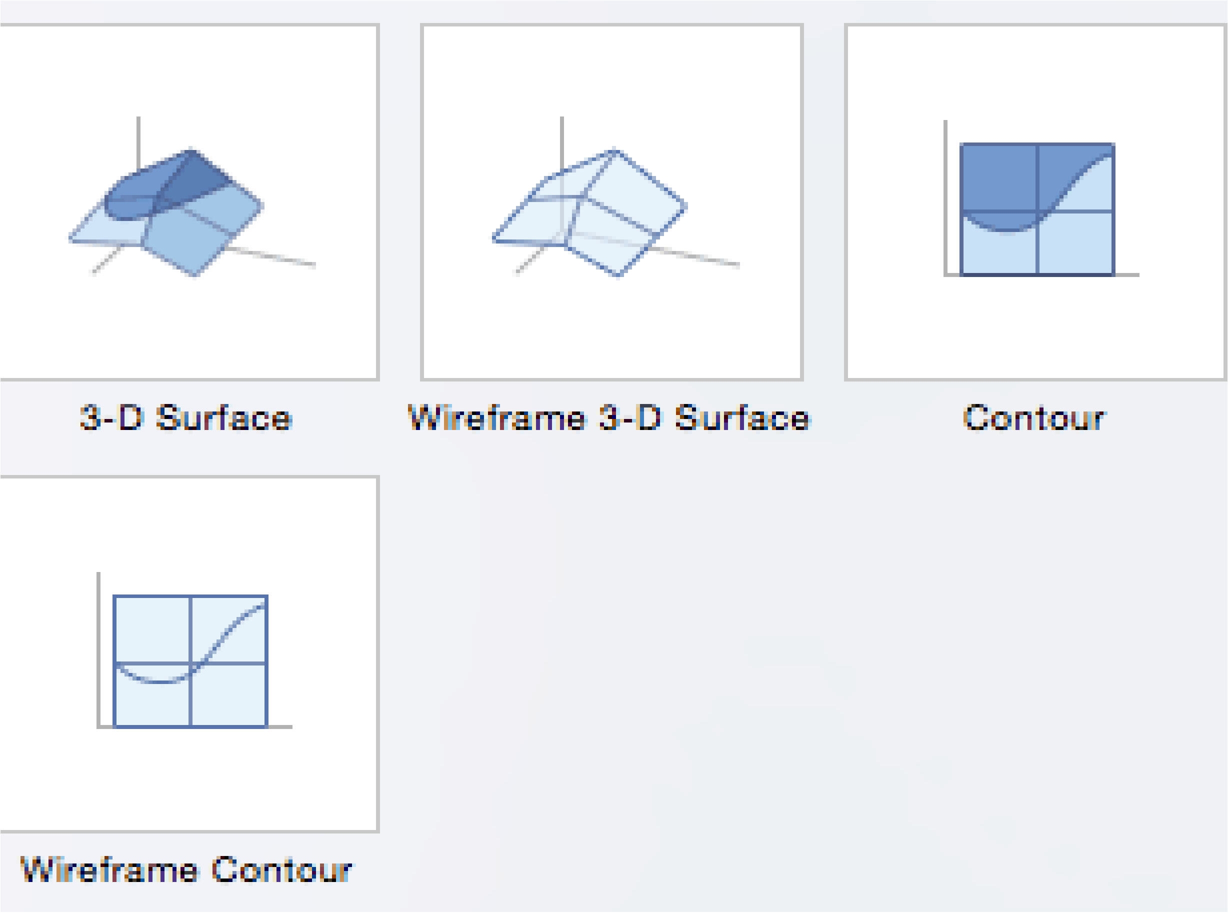


2. **Special charts:** Excel includes several charts suitable for presenting scientific statistical, and financial data. Scatter charts are used to present experimental results. Surface and cone charts are good for presenting 3-D and 2-D changes in data. Radar charts show data values about a single metric. Stock charts present values for between three and five series of data, including open, high, low, close, and volume trading information.

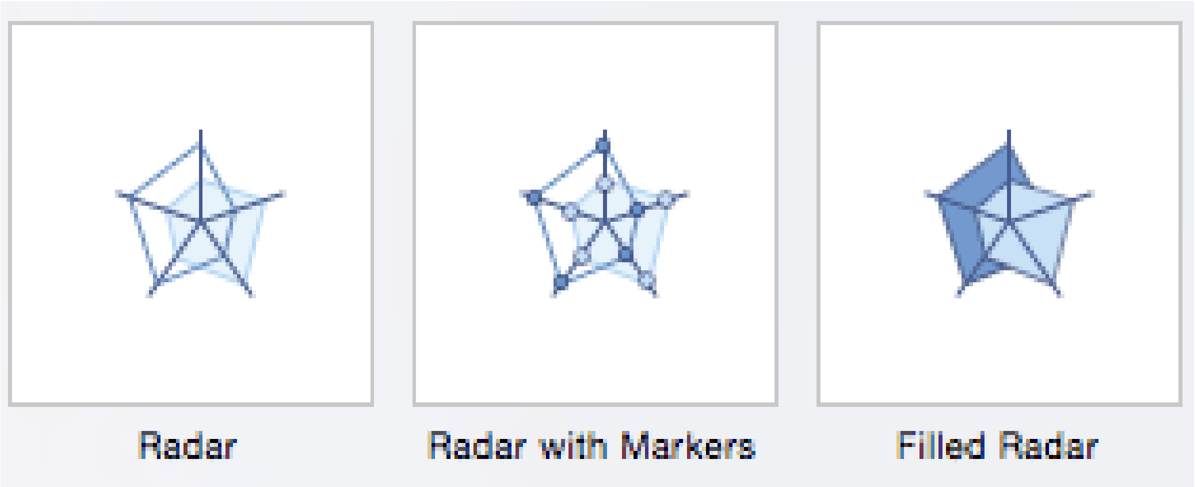
- **Scatter Charts:** Similar to line graphs, because they are useful for showing change in variables over time, scatter charts are used specifically to show how one variable affects another. (This is called correlation.) Note that bubble charts, a popular chart type, are categorized under scatter. There are seven scatter chart options: scatter, scatter with smooth lines and markers, scatter with smooth lines, scatter with straight lines and markers, scatter with straight lines, bubble, and 3-D bubble.


Scatter
Scatter with Smooth Lines and Markers
Scatter with Smooth Lines
Scatter with Straight Lines and Markers
Scatter with Straight Lines
Bubble
3-D Bubble

- **Surface:** Use a surface chart to represent data across a 3-D landscape. This additional plane makes them ideal for large data sets, those with more than two variables, or those with categories within a single variable. However, surface charts can be difficult to read, so make sure your audience is familiar with them. You can choose from 3-D surface, wireframe 3-D surface, contour, and wireframe contour.

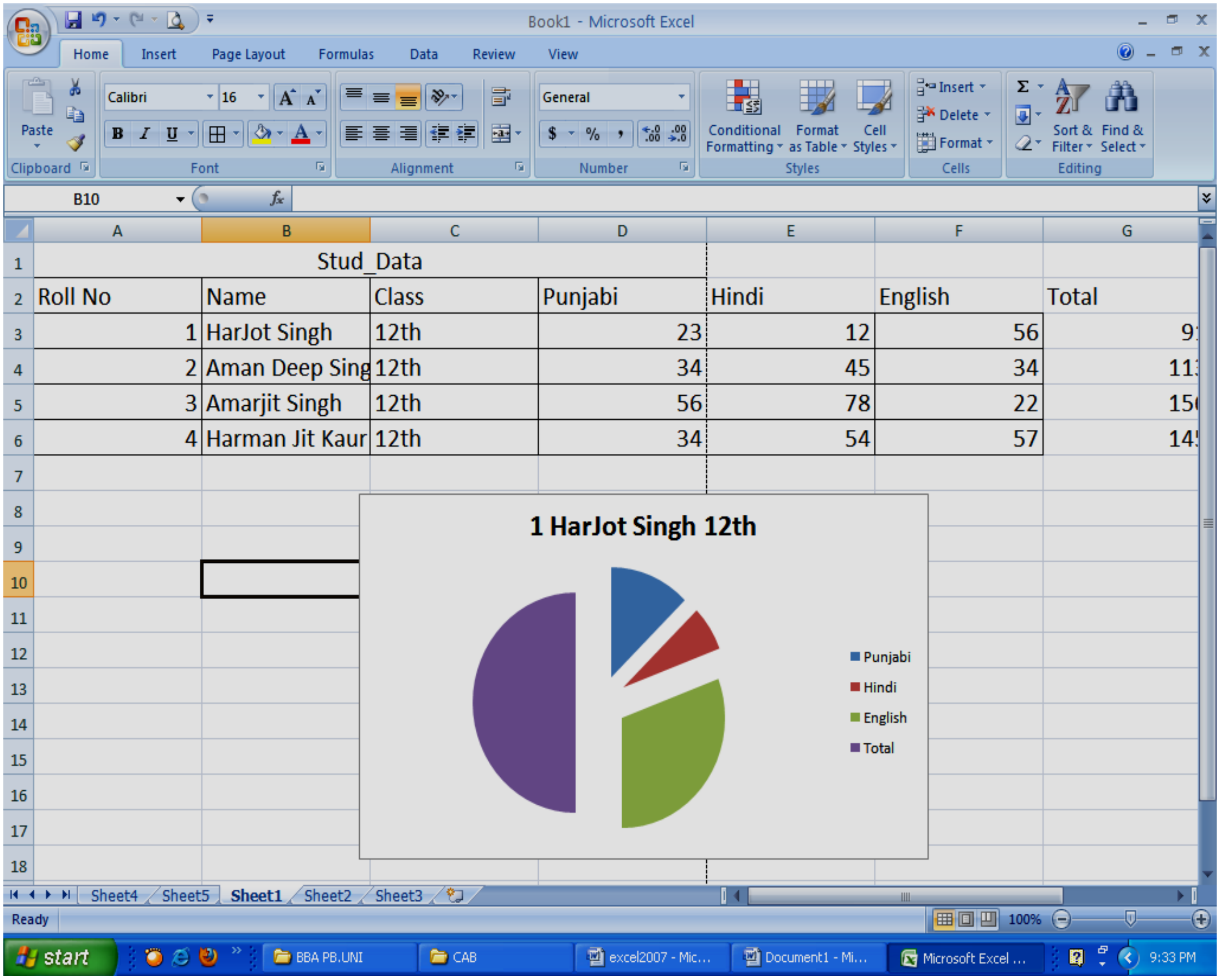


- **Radar:** When you want to display data from multiple variables about each other use a radar chart. All variables begin from the central point. The key with radar charts is that you are comparing all individual variables about each other — they are often used for comparing the strengths and weaknesses of different products or employees. There are three radar chart types: radar, radar with markers, and filled radar.



1. Pie Charts

Use pie charts to show the relationships between pieces of an entity. The implication is that the pie includes all or something. The pie chart isn't appropriate for illustrating some of anything, so if there's not an obvious "all" in the data you're charting, don't use a pie.



Stud_Data

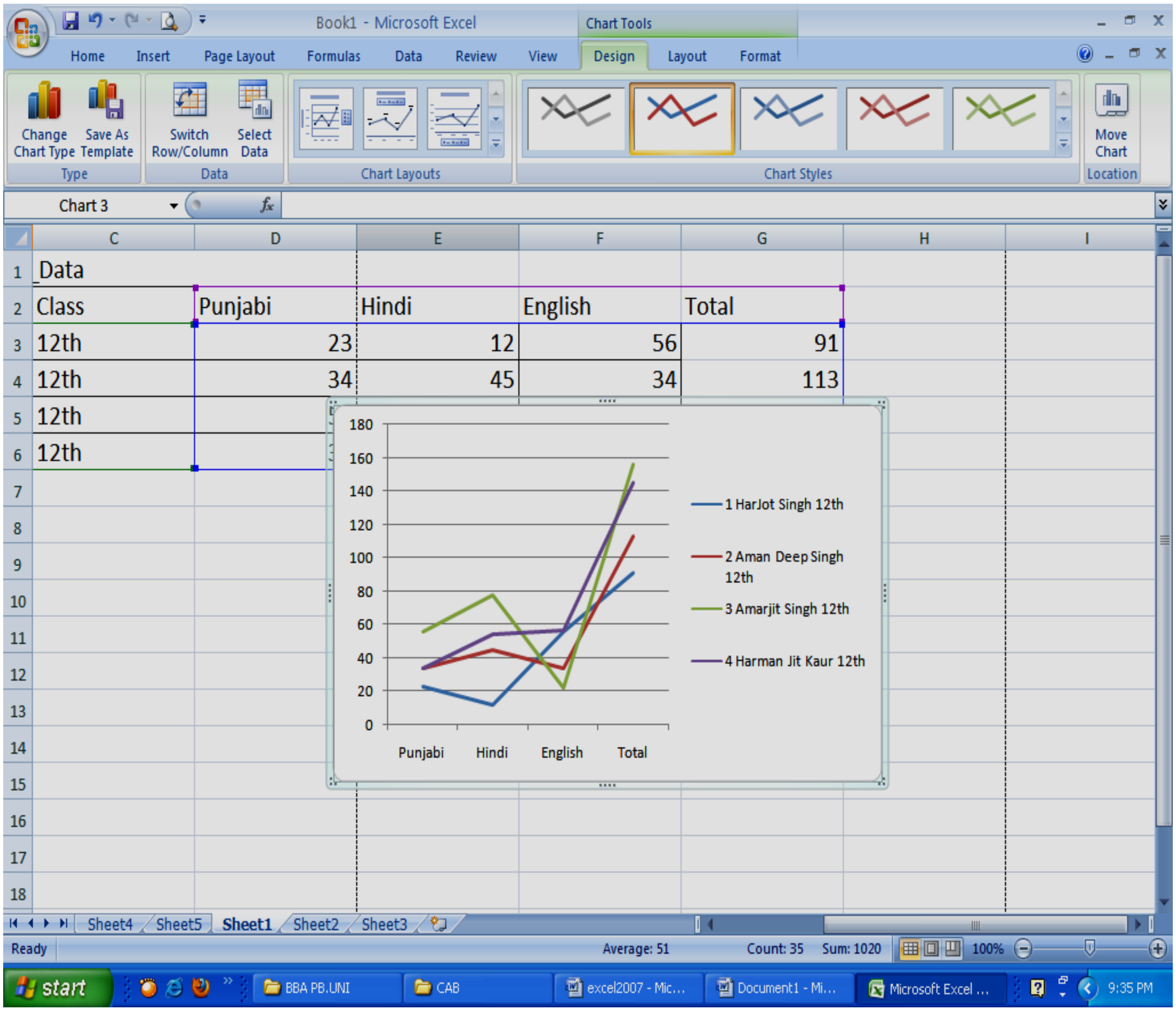
Roll No	Name	Class	Punjabi	Hindi	English	Total
1	HarJot Singh	12th	23	12	56	9
2	Aman Deep Sing	12th	34	45	34	11
3	Amarjit Singh	12th	56	78	22	15
4	Harman Jit Kaur	12th	34	54	57	14

Fig. 1: Pie Charts

A pie chart can only include one data series. If you select more than one data, series, Excel uses the first series and ignores all others. No error message appears, so you won't necessarily know that the chart doesn't show the data you intended to include unless you examine the chart carefully. When you create a pie chart, Excel totals the data points in the series and then divides the value of each data point into the series total to determine how large each data point's pie slice should be. Don't include a total from the worksheet as a data point; this doubles the total Excei calculates, resulting in a pie chart with one large slice (50 percent of the pie).

2. Line and Area Charts

The series chart shown in the figure is a line chart. In a 2-D version (as shown) or in a 3-D version that is sometimes called a ribbon chart. An area chart is a line chart with the area below the line filled. Line charts and area charts are typically used to show one or more variables (such as sales, income, or price) changing over time.



	C	D	E	F	G
1	Data				
2	Class	Punjabi	Hindi	English	Total
3	12th	23	12	56	91
4	12th	34	45	34	113
5	12th				
6	12th				

Average: 51 Count: 35 Sum: 1020

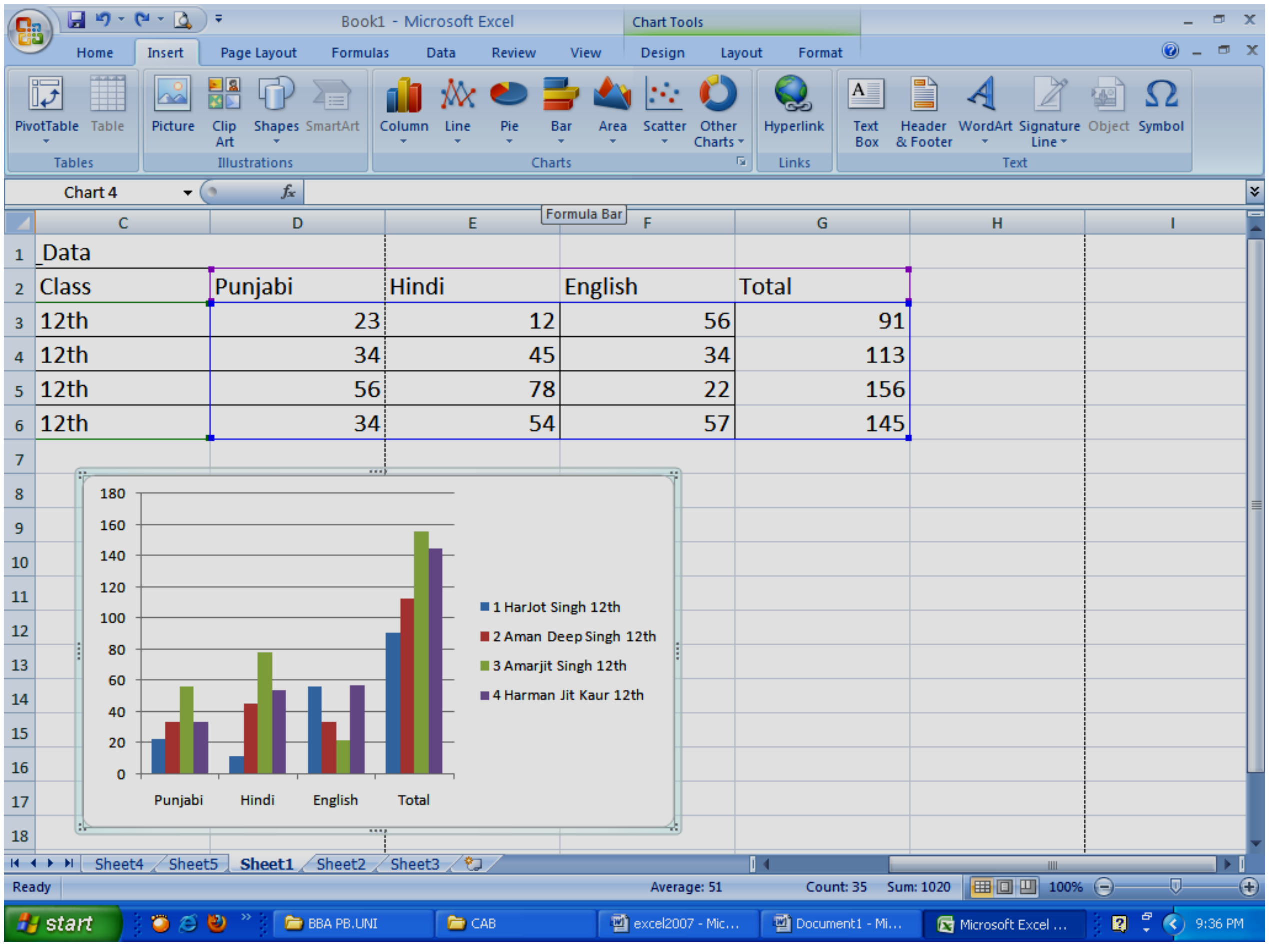
Fig. 2: Line and Area Charts

3. Column Chart

The figure shows the same information presented as a bar chart. The bars give added substance to the chart. In the line chart, what the reader notices is the trend up or down in each line and the gaps between the lines.

Line and area charts share a common layout. The horizontal line is called the X-axis, and the vertical line is the Y-axis (the same x- and y-axis you may have learned about in algebra or geometry class when plotting data points). In a bar chart, however, the axis is turned 90 degrees so that the x-axis is on the left side.

Excel can also combine columns with line or area charts and embellish line or column charts with 3-D effects. You can make the columns and lines on your charts into tubes, pyramids, cones, or cylinders; or transform regular bars into floating 3-D bars. Plotting data on two axes is also possible with column charts.



Data

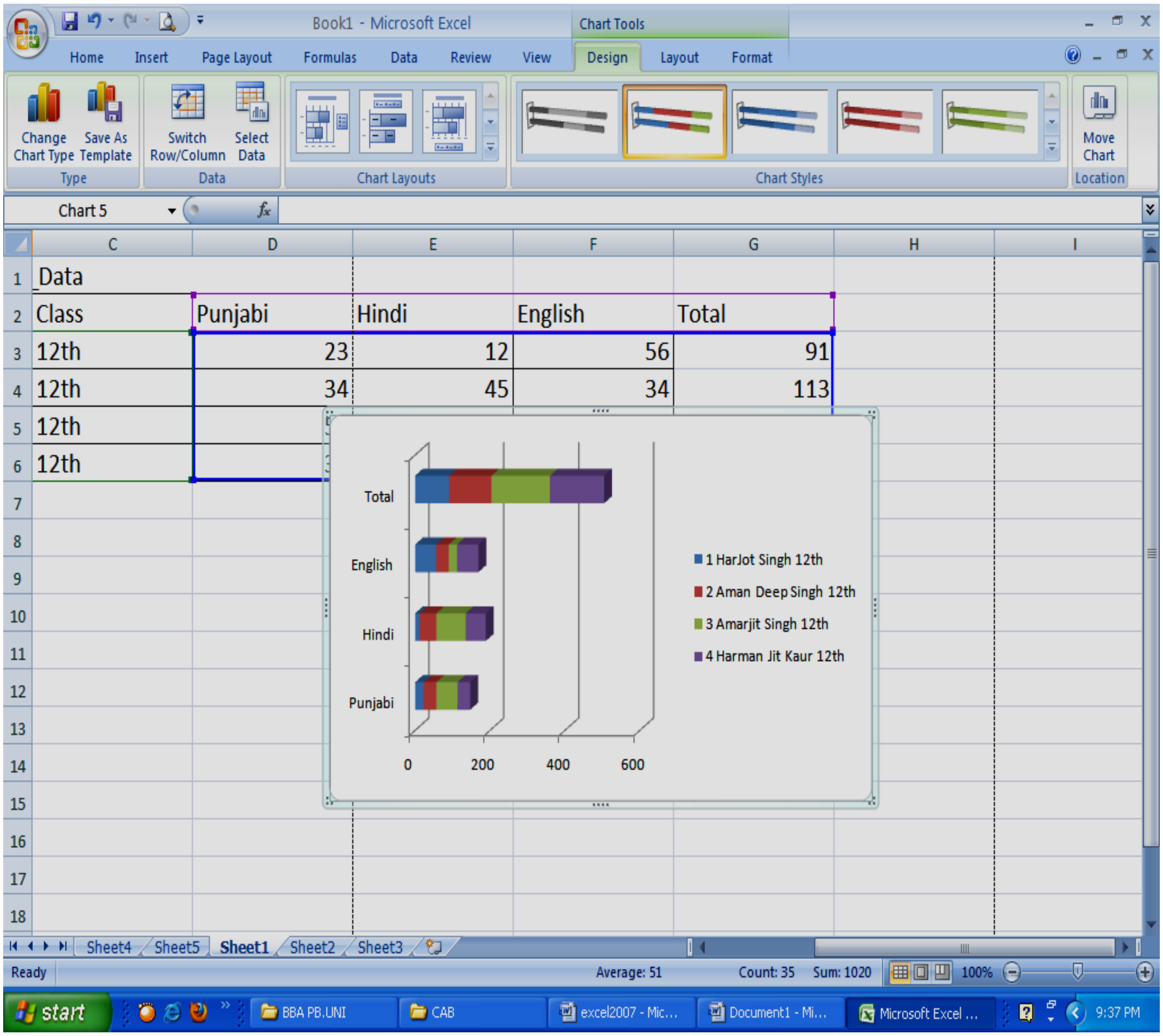
Class	Punjabi	Hindi	English	Total
12th	23	12	56	91
12th	34	45	34	113
12th	56	78	22	156
12th	34	54	57	145

Ready Average: 51 Count: 35 Sum: 1020 100%

Fig. 3: Column Chart

4. Bar Chart Variations

Column charts are the same as bar charts but with the X-axis at the bottom. There are three-dimensional varieties of bar and column charts, which add depth to the regular chart. Cylinders, cones, and pyramids are variations of a column chart.



	C	D	E	F	G
1	Data				
2	Class	Punjabi	Hindi	English	Total
3	12th	23	12	56	91
4	12th	34	45	34	113
5	12th				
6	12th				

Ready   Average: 51   Count: 35   Sum: 1020   100%

Fig. 4: Bar Chart Variations

Excel also offers another style of bar and column chart–the stacked chart. A stacked 3-D column chart, using the same data as Figure. In a stacked chart, parallel data points in each data series are stacked on top or to the right of each other. Stacking adds another dimension to the chart since it allows the user to compare sales between as well as within time periods providing a column chart and a pie chart for each period.

The 3-D charts have three axes. In a 3-D column chart, the X-axis is on the bottom. The vertical axis is the Z-axis; the Y-axis goes from front to back, providing the “third dimension” of depth in the chart. Don’t worry about memorizing which axis is which in each chart type; there are ways to know which is which when you’re creating or editing the chart.

3.6 APPLY CHART LAYOUT

Context tabs are tabs that only appear when you need them. Called Chart Tools, there are three chart context tabs: Design, Layout, and Format. The tabs become available when you create a new chart or when you click on a chart. You can use these tabs to customize your chart. You can determine what your chart displays by choosing a layout. For example, the layout you choose determines whether your chart displays a title, where the title displays, whether your chart has a legend, where the legend displays, whether the chart has axis labels, and so on. Excel provides several layouts from which you can choose.

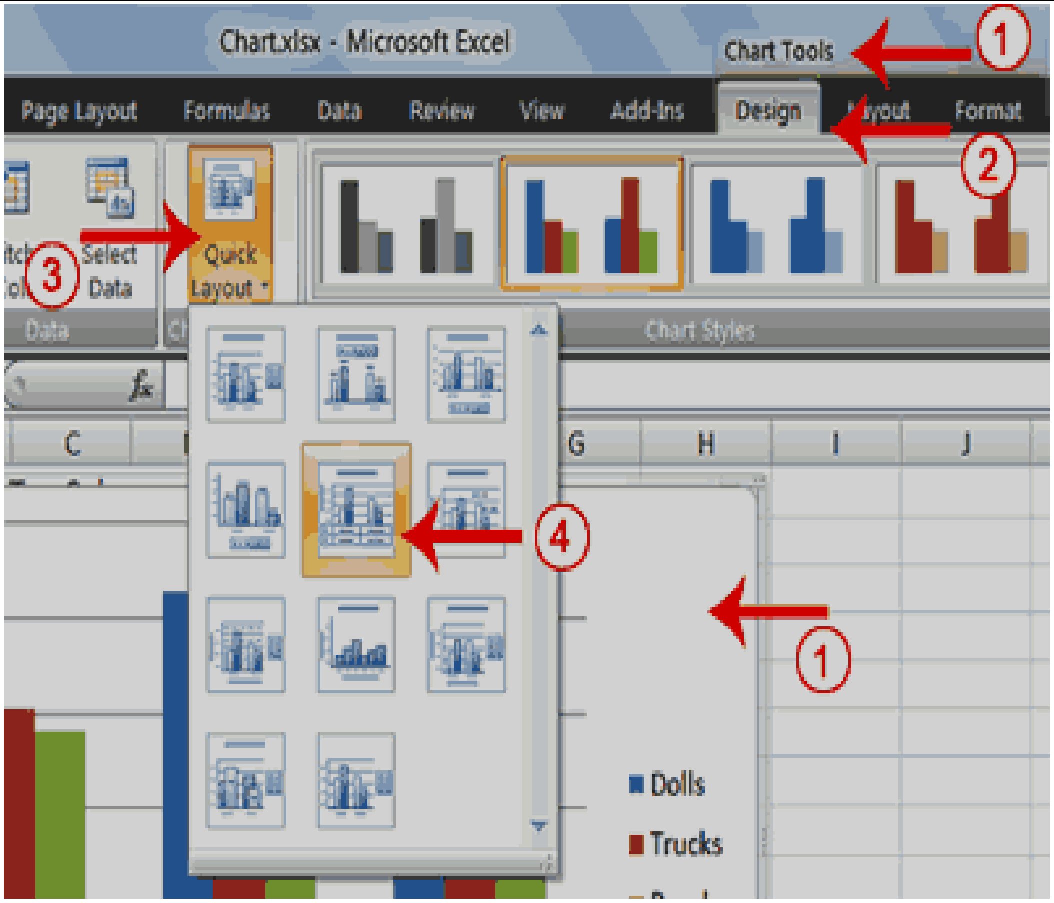


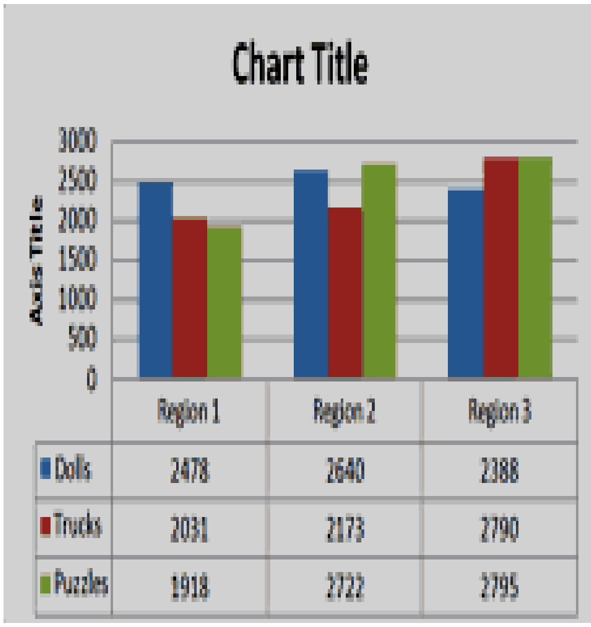
Fig Chart Layout

Steps to Apply a Chart Layout

1. Click your chart. The Chart Tools become available.
2. Choose the Design tab.
3. Click the Quick Layout button in the Chart Layout group. A list of chart layouts appears.
4. Click Layout. Excel applies the layout to your chart.

3.7 ADD LABELS

When you apply a layout, Excel may create areas where you can insert labels. You use labels to give your chart a title or to label your axes. When you applied layout, Excel created label areas for a title and the vertical axis



Before

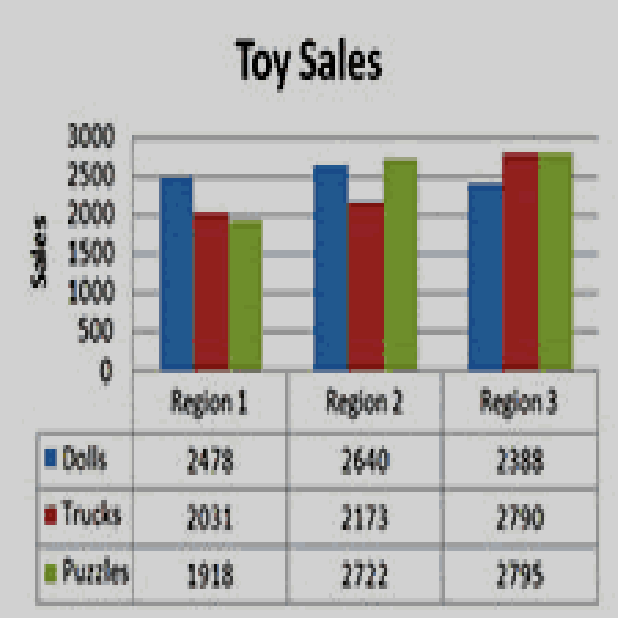


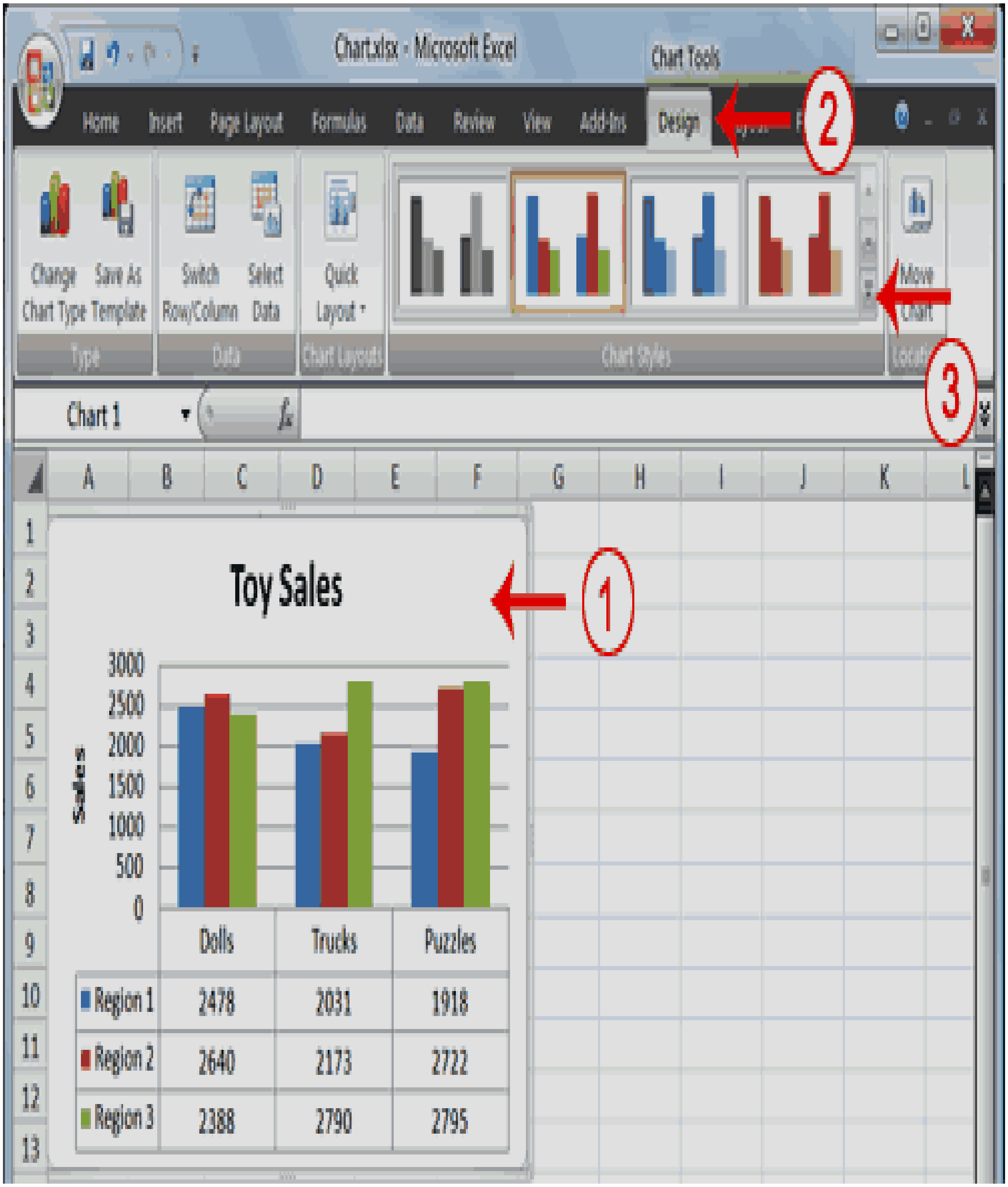
Fig After

Steps to add labels

1. Select Chart Title. Click on Chart Title and then place your cursor before the C in Chart and hold down the Shift key while you use the right arrow key to highlight the words Chart Title.
2. Type **Toy Sales**. Excel adds your title.
3. Select Axis Title. Click on Axis Title. Place your cursor before the A in Axis. Hold down the Shift key while you use the right arrow key to highlight the words, Axis Title.
4. Type Sales**.** Excel labels the axis.
5. Click anywhere on the chart to end your entry.

3.8 CHANGE THE STYLE OF A CHART

A style is a set of formatting options. You can use a style to change the color and format of your chart. Excel has several predefined styles that you can use. They are numbered from left to right, starting with 1, which is located in the upper-left corner.

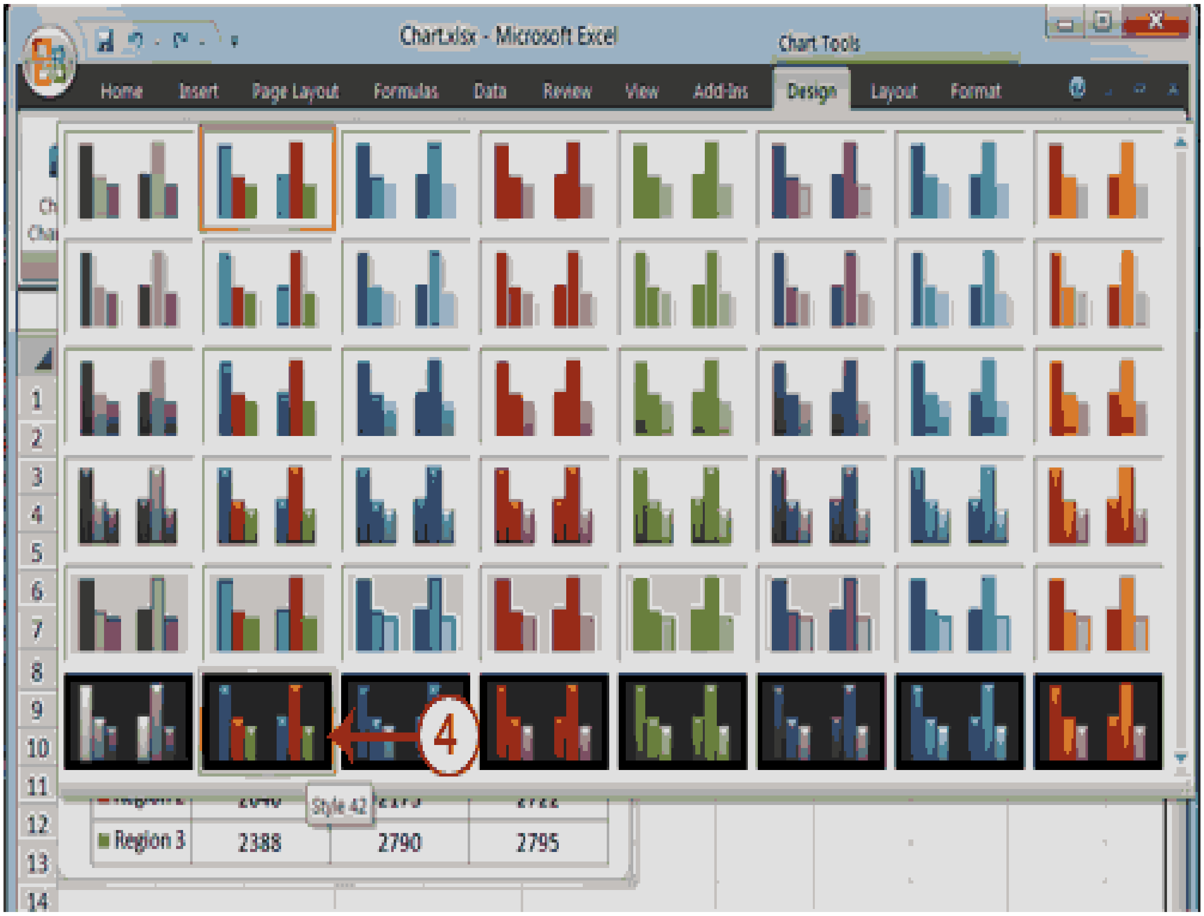


	Dolls	Trucks	Puzzles
🟦 Region 1	2478	2031	1918
🟥 Region 2	2640	2173	2722
🟩 Region 3	2388	2790	2795

Fig. Change the Style of a Chart

Steps to Change the Style of a Chart

1. Click your chart. The Chart Tools become available.
2. Choose the Design tab.
3. Click the More button in the Chart Styles group. The chart styles appear.



Fig

4. Click Style. Excel applies the style to your chart.

3.9 DATA PRESERVATION

Data preservation is the act of conserving and maintaining both the safety and integrity of data. Preservation is done through formal activities that are governed by policies, regulations, and strategies directed towards protecting and prolonging the existence and authenticity of data. Data preservation refers to maintaining access to data and files over time. For data to be preserved, at minimum, it must be stored in a secure location, stored across multiple locations, and saved in file formats that will likely have the greatest utility in the future. Data preservation provides the usability of data beyond the lifetime of the research activity that generated them.

Definition

Data preservation consists of a series of managed activities necessary to ensure continued stability and access to data for as long as necessary. For data to be preserved, at minimum, it must be stored in a secure location, stored across multiple locations, and saved in open file formats that will likely have the greatest utility in the future. Part of the preservation process can include depositing data in an institutional, discipline-specific, or generalist data repository, all of which allow for publication and preservation.

- The new NIH Data Management and Sharing Policy requires data to be preserved and shared, so medical researcher submits their COVID data to the National COVID Cohort Collaborative (N3C), as listed in the Open Domain-Specific Data Sharing Repository.
- There are many different ways to preserve digital information, including text, photos, audio, and video. Some strategies to preserve digital information include: Refreshing: transferring data to the same format. An example would be the transfer of music from an old CD-ROM to a new CD-ROM.

Digital information is an important source in our knowledge economy, valuable for research and education, science and the humanities, creative and cultural activities, and public policy. New high-throughput instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. These data are used by decision-makers to improve the quality of life of citizens. Moreover, researchers are employing sophisticated technologies to analyze these data to address questions that were unapproachable just a few years ago. Digital technologies have fostered a new world of research characterized by immense datasets, unprecedented levels of openness among researchers, and new connections among researchers, policymakers, and the public domain. Different types of threats are:

1. Users may be unable to understand or use the data,
2. Lack of sustainable hardware, software, or support of the computer environment may make the information inaccessible,
3. Evidence may be lost because the origin and authenticity of the data may be uncertain,
4. Access and use restrictions (e.g., Digital Rights Management) may not be respected in the future,
5. Loss of ability to identify the location of data,
6. The current custodian of the data, whether an organization or project, may cease to exist at some point in the future.

➢ Importance of Data Preservation

Preservation helps protect you from hardware obsolescence. You never want to find yourself in a situation where all of your data is saved on unsupported hardware! Always migrate to new hardware formats so that your data will be available long-term. The most responsible way to preserve your data is to turn it over to a responsible custodian such as a data repository. When possible, try to preserve research data in a repository that provides *data curation services*, not just preservation services. Curated data is more valuable, easier to reuse, easier to locate, and more highly cited. Many data repositories have requirements for deposit - they may only accept certain types of data and have file size limits.

3.10 DATA PRESERVING VS. STORING DATA

Preserving is different from storing and backing up data files while your research is still ongoing. The latter typically involves mutable data; the former concerns data (or milestone versions of data) that are 'frozen' and not in active use. Long-term preservation requires appropriate actions to prevent data from becoming unavailable and unusable over time, for example, because of:

- Outdated software or hardware
- Storage media degradation
- A lack of sufficient descriptive and contextual information to keep data understandable

In other words, data preservation involves more than simply not deleting the data files created and stored. Maintaining data in a usable form for the longer term takes effort and has a considerable cost. Selecting which (parts of) data to keep, and for how long, is, therefore, an essential component of data preservation.

As a researcher, you have a key role in deciding what to retain and what not, as you know your data best. Such decisions may depend on factors such as:

- The type of data involved
- The norms in your discipline
- Whether you are keeping data for potential future reuse, verification, or other purposes. Depending on the purpose, you may need to keep the raw data or data in a more processed form.

3.11 DATA PRESERVATION VS. RETENTION OF DATA

- Data retention is a central component of records management and information governance. Retention refers to the storing of data to meet regulatory and recordkeeping obligations
- The preservation is related to the safekeeping of electronically stored information (ESI) for some anticipated legal matter. In other words, data retention is a proactive ongoing process.
- Retention is usually a mandated requirement for researchers - it's the task that ensures that a bare minimum of data will remain available in some format.
- Preservation refers to having an active plan to ensure that when you do need to access your old data, it's readily available and can be easily accessed and manipulated by whoever needs it. When making a plan for data preservation you should include activities such as:
- Transferring data from older storage formats to newer ones. This will ensure that the technology required to access your data is still available.
- Transferring data from older file formats to newer ones. This will ensure that your data can still be opened by current software applications.
- Having multiple copies of the data in different locations. This will ensure that your data is not lost in an unexpected event, such as theft, flood, or fire.
- Ensuring your data is well documented, such as making notes on software used for creating codebooks as outlined in the Documentation and Description section of this guide. This will ensure that when you come back to access your data, you'll be able to remember what it all means.

3.12 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions (Define the followings)

Q1. data processing

Q2. Chart Layout

Q3. Add Labels

Q4. Change the Style of a Chart

Q5. Data Preserving

Q6. Data Preserving vs. Storing Data

Q7. Pie chart

Q8. Bar chart

Q9. Line and area Chart

Q10. Surface chart

Q11. Radar chart

B. Long Answer Type Questions

Q1. Define data processing. What are the various stages of data processing?

Q2. What is the purpose of data processing?

Q3. What is the significance of data processing in the research?

Q4. Explain the various rules for data processing.

Q5. What is the role of Excel in the presentation of data processing?

Q6. What is chart layout? How it can be applied in Excel?

Q7. What is the utility of the style of the chart in data processing? How it can be applied?

Q8. Define data preservation. What are the various challenges and threats to the digital preservation of data?

Q9. Explain data preservation Vs. storage of data.

3.13 SUGGESTED READINGS

- Abebe, J. Daniels, J.W. Mckean, "Statistics and Data Analysis".
- Clarke, G.M. & Cooke, D., "A Basic Course in Statistics", Arnold.
- David M. Lane, "Introduction to Statistics".
- S.C. Gupta and V.K. Kapoor, "Fundamentals of Mathematical Statistics", SultanChand & Sons, New Delhi.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 4: MEASURES OF CENTRAL TENDENCY

STRUCTURE

4.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Know the meaning of average
- Features of good measure of average
- Find different types of averages for various types of data
- Understand the relation that exists between different types of Averages
- Know merits and limitations of each type of average
- To calculate mean, mode and medium for different series

4.1 INTRODUCTION

We can say that modern age is the age of Statistics. There is no field in the modern life in which statistics is not used. Whether it is Business, Economics, Education. Government Planning or any other field of our life, statistics is used everywhere. Business manager use statistics for business decision making, Economists use statistics for economic planning, Investors use statistics for future forecasting and so on. There are many techniques in statistics that helps us in all these purposes. Average or Central Tendency is one such technique that is widely used in statistics. This technique is used almost in every walk of the life.

4.2 MEANING OF AVERAGE OR CENTRAL TENDENCY

Average or Central tendency is the most used tool of statistics. This is the tool without which statistics is incomplete. In simple words we can say that Average is the single value which is capable of representing its series. It is the value around which other values in the series move. We can define Average as the single typical value of the series which represent whole data of the series. Following is the popular definition of average:

As per **Croxton and Cowden** "An average is a single value within the range of data that is used to represent all values in the series. Since an average is somewhere within the range of the data, it is also called a measure of Central Value".

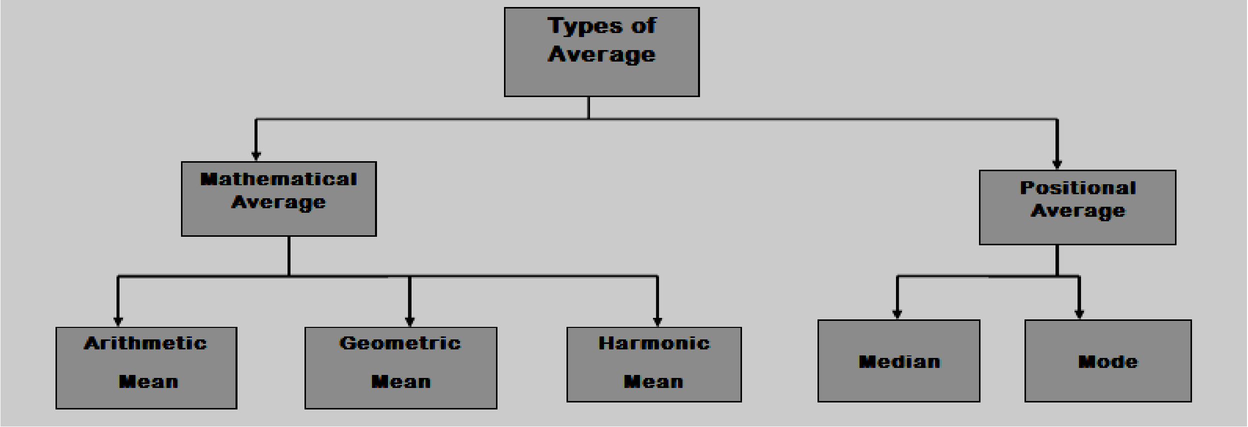
4.3 OBJECTIVES AND FUNCTIONS OF AVERAGE

1. **Single Value representing whole Data:** In statistics data can be shown with the help of tables and diagrams. But some time data is very larger and it is not easy to present in table or graph. So, we want to represent that data in summarised form. Average helps us to represent data in summarised form. For example, that data of national income of India is very large but when we calculate per capita income it gives us idea of the national income.
2. **To Help in Comparison**: In case we want to compare two different series of data, it is very difficult to compare. There are many difficulties like number of items in the series may be different. In such case average helps us in making the comparison. For example, if we want to compare income of people living in different countries like India and Pakistan, we can do so by calculating per capita income which is a form of average.
3. **Draw conclusion about Universe from Sample:** This is one of the important functions of average. If we take the average of a sample, we can draw certain conclusion about the universe from such Average. For example, mean of a sample is representative of its universe.
4. **Base of other Statistical Methods:** There are many Statistical Techniques that are based on average. If we don't have an idea about the average, we cannot apply those techniques. For example, Dispersion, Skewness, Index Number are based on average.
5. **Base of Decision Making**: Whenever we have to make certain decision, average plays very crucial role in the decision making. From the average we could have idea about the data and on the basis of that information we can take decision. For example, a company can take decision regarding its sales on the basis of average yearly sales of past few years.

6. **Precise Relationship**: Average helps us to find out if there is precise relation between two variables or two items. It also removes the biasness of the person making analysis. For example, if you say that Rajesh is more intelligent than Ravi it is only our personal observation and does not make any precise relation. If we compare the average marks of both the students, we could have a precise relation.
7. **Helpful in Policy Formulation**: Average helps the government in formulation of the policy. Whenever government has to formulate economic policy they consider various averages like per capita income, average growth rate etc.

4.4 MEASURES OF CENTRAL TENDENCY

There are many methods through which we can calculate average or central tendency. We can divide these methods into two categories that are Algebraic Method and Positional Average. Algebraic methods are those in which the value of average depends upon the mathematical formula used in the average. The mathematical average can further be divided into three categories that are Arithmetic Mean, Geometric Mean and Harmonic Mean. On the other hand positional average are those average which are not based on the mathematical formula used in calculation of average rather these depends upon the position of the variable in the series. As these depends upon the position of the variable, these averages are not affected by the extreme values in the data. Following chart shows different types of averages.



4.5 ARITHMETIC MEAN

It is the most popular and most common measure of average. It is so popular that for a common man the two terms Arithmetic Mean and Average are one and the same thing. However, in reality these two terms are not same and arithmetic mean is just one measure of the average. We can define the arithmetic mean as:

"The value obtained by dividing sum of observations with the number of observations".

So arithmetic mean is very easy to calculate, what we have to do is just add up the value of all the items given in the data and then we have to divide that total with the number of items in the data. Arithmetic mean is represented by symbol $\mathrm{A.M.}$ or $\overline{\times}$

4.5.1 Arithmetic Mean in case of Individual Series

Individual series are those series in which all the items of the data are listed individually. There are two methods of finding arithmetic mean in the individual series. These two methods are Direct method and Shortcut Method.

1. **Direct Method** According to this method calculation of mean is very simple and as discussed above, we have to just add the items and then divide it by number of items. Following are the steps in calculation of mean by direct method:
   1. Suppose our various items of the data are X_1, X_2, X_3 ………………….. X_n
   2. Add all the values of the series and find ∑X.
   3. Find out the number of items in the series denoted by n.
   4. Calculate arithmetic mean dividing sum value of observation with the number of observations using following formula:

$$\overline{\times} = \frac{\mathbf{X1 + X2 + X3 + -------Xn}}{\mathbf{N}} = \frac{\sum \mathbf{X}}{\mathbf{N}}$$

Where $\overline{\times}$= Mean

N = Number of items

∑X = Sum of observation

Example 1. The daily income of 10 families is a as given below (in rupees):

$\mathbf{130, 141, 147, 154, 123, 134, 137, 151, 153, 147}$

Find the arithmetic mean by direct method.

Solution: Computation of Arithmetic Mean

Serial No.	Daily Income (in Rs.) X

1	130
2	141
3	147
4	154
5	123
6	134
7	137
8	151
9	153
10	147
$\mathrm{N} = 10$	$\sum \mathrm{X} = 1417$

$$\text{A. M.,}\qquad \overline{\times}= \frac{\mathrm{X1} + \mathrm{X2} + \mathrm{X3} + \;-------\mathrm{XN}}{\mathrm{N}} = \frac{\sum \mathrm{X}}{\mathrm{N}} = \frac{1417}{10} = \mathrm{Rs.}\,141.7$$

2. **Short Cut Method:** Normally this method is used when the value of items is very large and it is difficult to make calculations. Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. This method is also known as is assumed mean method. Following are the steps of this method:
    1. Suppose our various items of the data are X_1, X_2, X_3 ………………….. X_n
    2. Take any value as assumed mean represented by 'A'. This value may be any value among data or any other value even if that is not presented in data.
    3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are representing as 'dx'
    4. Find sum of the deviations represented by ∑dx.
    5. Find out the number of items in the series denoted by n.
    6. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{\times}= \mathbf{A} + \frac{\sum \mathbf{dx}}{\mathbf{N}}$$

Where $\overline{\times}$= Mean

A = Assumed Mean

N = Number of items

∑dx = Sum of deviations

Example 2. Calculate A. M. by short - cut method for the following data

R. No.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Marks	**50**	**60**	**65**	**88**	**68**	**70**	**83**	**45**	**53**	**58**

Solution: Let assumed Mean (A) be 60

R. No.	Marks (X)	$dx = X - A$
1	50	-10
2	60	0
3	65	5
4	88	28
5	68	8
6	70	10
7	83	23
8	45	-15
9	53	-7
10	58	-2
N = 10		$\sum dx = 40$

As $\overline{X} = A + \frac{\sum dx}{N}$

⇒ $\overline{X} = 60 + \frac{40}{10} \quad = 60 + 4$

⇒ $\overline{X} = 64$ Marks

4.5.2 Arithmetic Mean in case of Discrete Series

In individual series if any value is repeated that is shown repeatedly in the series. It makes series lengthy and make calculation difficult. In case of discrete series, instead of repeatedly showing the items we just group those items and the number of time that item is repeated is shown as frequency. In case of discrete series, we can calculate Arithmetic mean. By using Direct Method and Shortcut Method.

1. **Direct Method:** In indirect method we multiply the value of items (X) with their respective frequency (f) to find out the product item (fX). Then we take up sun of the product and divide it with the number of items. Following are the steps
   1. Multiply the value of items (X) with their respective frequency (f) to find out the the product item (fX)
   2. Add up the product so calculated to find ∑fX.
   3. Find out the number of items in the series denoted by n.

4. Calculate arithmetic mean dividing sum of the product with the number of observations using following formula:

$$\overline{X} = \ = \frac{\sum \mathbf{fX}}{\mathbf{N}}$$

Where $\overline{X}$= Mean

N = Number of items

∑fX = Sum of product of observations.

Example 3. Find the average income

Daily Income (in rupees)	**200**	**500**	**600**	**750**	**800**
No. of Workers	**2**	**1**	**4**	**2**	**1**

Solution:

Daily Income (Rs.) X	No. of Workers Frequency (f)	fX
200	2	400
500	1	500
600	4	2400
750	2	1500
800	1	800
	$\sum$ f = 10	$\sum$ fX = 5600

∴ Average Income $\overline{X} = \frac{\sum fX}{\sum f}$

$$= \frac{5600}{10} \quad = \text{Rs.}\,560$$

2. **Short Cut Method:** Under this method we take one value as mean which is known as assumed mean and deviations are calculated from this as you mean. Then average is calculated using assumed mean. Following are the steps of this method:
   1. Suppose our items of the data are 'X' and its corresponding frequency is 'f'.
   2. Take any value as assumed mean represented by 'A'.
   3. Find out deviations of items from assumed mean. For that deduct Assumed value from each value of the data. These deviations are representing as 'dx'
   4. Multiply the values of dx with corresponding frequency to find out product denoted by fdx

5. Find sum of the product so calculated represented by ∑fdx.
6. Find out the number of items in the series denoted by n.
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{\mathrm{X}} = \mathbf{A} + \frac{\sum \mathbf{fdx}}{\mathbf{N}}$$

Where $\overline{\mathrm{X}}$= Mean, A = Assumed Mean, N = Number of items

∑fdx = Sum of product of deviation with frequency.

Example 4. From the following data find out the mean height of the students.

Height (in cms.)	**154**	**155**	**156**	**157**	**158**	**159**	**160**	**161**	**162**	**163**
No. of Students	**1**	**6**	**10**	**22**	**21**	**17**	**14**	**5**	**3**	**1**

Solution: Let the Assumed Mean (A) be 150

Height in cms. X	No. of students f	$dX = (X - A)$ $= X - 150$	fdX
154	1	4	4
155	6	5	30
156	10	6	60
157	22	7	154
158	21	8	168
159	17	9	153
160	14	10	140
161	5	11	55
162	3	12	36
163	1	13	13
	$\sum f = 100$		$\sum fdX = 813$

Applying the formula

$$\overline{\mathrm{X}} = \mathrm{A} + \frac{\sum \mathrm{fdX}}{\sum \mathrm{f}}$$

We get

$$\overline{\mathrm{X}} = 150 + \frac{813}{100}$$
$$= 150 + 8.13$$
$$= 158.13$$

∴ Mean Height $= 158.13$ cm

4.5.3 Arithmetic Mean in case of Continuous Series

Continuous series is also known as Grouped Frequency Series. Under this series the values of the observation are grouped in various classes with some upper and lower limit. For example, classes like 10-20, 20-30, 30-40 and so on. In the class 10-20 lower limit is 10 and upper limit is 20. So, all the observations having values between 10 and 20 are put in this class interval. Similar procedure is adopted for all class intervals. The procedure of calculating Arithmetic Mean is continuous series is just like discrete series except that instead of taking values of observations we take mid value of the class interval. The mid value is represented by 'm' and is calculated using following formula:

$$\mathbf{m = \frac{Lower\ Limit + Upper\ Limit}{2}}$$

1. **Direct Method:** In indirect method we multiply the mid values (m) with their respective frequency (f) to find out the product item (fm). Then we take up sun of the product and divide it with the number of items. Following are the steps
    1. Multiply the mid values (m) with their respective frequency (f) to find out the the product item (fm)
    2. Add up the product so calculated to find ∑fm.
    3. Find out the number of items in the series denoted by n.
    4. Calculate arithmetic mean by dividing sum of the product with the number of observations using following formula:

$$\overline{\times} = \ = \frac{\sum \mathbf{fm}}{\mathbf{N}}$$

Where $\overline{\times}$= Mean

N = Number of items

∑fm = Sum of product of observations of mean and frequencies.

Example 5. Calculate the arithmetic mean of the following data:

Class Intervals (C. I.)	**100 − 200**	**200 − 300**	**300 − 400**	**400 − 500**	**500 − 600**	**600 − 700**
f	**4**	**7**	**16**	**20**	**15**	**8**

Solution:

Class Intervals C. I.	Mid Value m	Frequency f	fm
$100 - 200$	150	4	600
$200 - 300$	250	7	1750
$300 - 400$	350	16	5600
$400 - 500$	450	20	9000
$500 - 600$	550	15	8250
$600 - 700$	650	8	5200
		$\sum f = 70$	$\sum fm = 30{,}400$

As $\overline{x} = \frac{\sum fm}{\sum f}$

∴ $\overline{x} = \frac{30{,}400}{70}$

$= 434.3$

2. **Short Cut Method:** This method of mean is almost similar to calculation in the discrete series but here the assumed mean is selected and then the deviation is taken from mid value of the observations. Following are the steps of this method:

1. Calculate the Mid Values of the series represented by ‘m’.
2. Take any value as assumed mean represented by ‘A’.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as ‘dm’
4. Multiply the values of dm with corresponding frequency to find out product denoted by fdm
5. Find sum of the product so calculated represented by ∑fdm.
6. Find out the number of items in the series denoted by n.
7. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{x} = \mathbf{A} + \frac{\sum \mathbf{fdm}}{\mathbf{N}}$$

Where $\overline{x}$ = Mean

A = Assumed Mean

N = Number of items

∑fdm = Sum of product of deviation from mid values with frequency.

Example 6. Calculate the mean from the following data

Daily Wages (Rs.)	0 − 100	100 − 200	200 − 300	300 − 400	400 − 500	500 − 600	600 − 700	700 − 800	800 − 900
No. of Workers	1	4	10	22	30	35	10	7	1

Solution: Let the assumed mean, A = 150

Daily Wages (Rs.) C.I.	No. of Workers f	Mid Value m	$dm = m - A$ $(m - 150)$	fdm
0 − 100	1	50	−100	−100
100 − 200	4	150	0	0
200 − 300	10	250	100	1000
300 − 400	22	350	200	4400
400 − 500	30	450	300	9000
500 − 600	35	550	400	14,000
600 − 700	10	650	500	5000
700 − 800	7	750	600	4200
800 − 900	1	850	700	700
	$\sum f = 120$			$\sum fdm = 38{,}200$

As $\overline{X} = A + \frac{\sum fdm}{\sum f}$

$= 150 + \frac{38{,}200}{120} \qquad = 150 + 318.33 = 468.33$

⇒ $\overline{X} = 468.33$

3. **Step Deviation Method:** Step Deviation method is the most frequently used method of finding Arithmetic Mean in case of continuous series. This method is normally used when the class interval of the various classes is same. This method makes the process of calculation simple. Following are the steps of this method:

1. Calculate the Mid Values of the series represented by ‘m’.
2. Take any value as assumed mean represented by ‘A’.
3. Find out deviations of items from assumed mean. For that deduct Assumed value from mid values of the data. These deviations are representing as ‘dm’.
4. Find out if all the values are divisible by some common factor ‘C’ and divide all the deviations with such common factor to find out dm’ which is dm/c

5. Multiply the values of dm’ with corresponding frequency to find out product denoted by fdm’
6. Find sum of the product so calculated represented by ∑fdm’.
7. Find out the number of items in the series denoted by n.
8. Calculate arithmetic mean dividing sum deviations of the observation with the number of observations using following formula:

$$\overline{x} = \mathbf{A} + \frac{\sum \mathbf{fdm'}}{\sum \mathbf{f}} \times \mathbf{C}$$

Where $\overline{x}$= Mean

A = Assumed Mean

N = Number of items

C = Common Factor

∑fdm’ = Sum of product of deviation after dividing with common factors and multiplying it with frequency.

Example 7. Use step deviation method to find $\overline{x}$ for the data given below:

Income (Rs.)	**1000 – 2000**	**2000 – 3000**	**3000 – 4000**	**4000 – 5000**	**5000 – 6000**	**6000 – 7000**
No. of Persons	**4**	**7**	**16**	**20**	**15**	**8**

Solution: Let the assumed mean A = 4500

Income (Rs.) C. I.	No of Persons f	Mid Value m	$dm = m - A$ $= (m - 4500)$	$dm' = \frac{dm}{C}$ $C = 1000$	fdm'
1000 – 2000	4	1500	−3000	−3	−12
2000 – 3000	7	2500	−2000	−2	−14
3000 – 4000	16	3500	−1000	−1	−16
4000 – 5000	20	4500	0	0	0
5000 – 6000	15	5500	1000	1	15
6000 – 7000	8	6500	2000	2	16
	$\sum f = 70$				$\sum fdm' = -11$

As $\qquad \overline{x} = A + \frac{\sum fdm'}{\sum f} \times C$

∴ $\qquad \overline{x} = 4500 + \frac{(-11)}{70} \times 1000$

$= 4500 - \frac{1100}{7} \qquad = 4500 - 157.14 \qquad = 4342.86$

$\overline{x} = 4342.86$

Other Special case of Continuous Series

4.5.4 Arithmetic Mean in case of Cumulative Frequency Series:

The normal continuous series give frequency of the particular class. However, in case of cumulative frequency series, it does not give frequency of particular class rather it gives the total of frequency including the frequency of preceding classes. Cumulative frequency series may be of two types, that are 'less than' type and 'more than' type. For calculating Arithmetic mean in cumulative frequency series, we convert such series into the normal frequency series and then apply the same method as in case of normal series.

Less than Cumulative Frequency Distribution

Example 8. Find the mean for the following frequency distribution:

Marks Less Than	**10**	**20**	**30**	**40**	**50**	**60**
No. of Students	**5**	**15**	**40**	**70**	**90**	**100**

Solution: Convert the given data into exclusive series:

Marks C. I.	No. of Students f	Mid Value m	$dm = m - A$ $A = 25$	$dm' = \frac{dm}{C}$ $C = 10$	fdm'
$0 - 10$	5	5	-20	-2	-10
$10 - 20$	$15 - 5 = 10$	15	-10	-1	-10
$20 - 30$	$40 - 15 = 25$	25	0	0	0
$30 - 40$	$70 - 40 = 30$	35	10	1	30
$40 - 50$	$90 - 70 = 20$	45	20	2	40
$50 - 60$	$100 - 90 = 10$	55	30	3	30
	$\sum f = 100$				$\sum fdm' = 80$

As $\overline{X} = A + \frac{\sum fdm'}{\sum f} \times C$

$\Rightarrow$ $\overline{X} = 25 + \frac{80}{100} \times 10 = 33$

$\Rightarrow$ $\overline{X} = 33$

More Than Cumulative Frequency Distribution

Example 9. Find the mean for the following frequency distribution

Marks More Than	**0**	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**
No. of Students	**80**	**77**	**72**	**65**	**55**	**43**	**28**	**16**	**10**	**8**

Solution: Convert the given data into exclusive series

Marks C. I.	No. of Students f	Mid Value	$dm = m - A$ $A = 55$	$dm' = \frac{dm}{C}$ $C = 10$	fdm'
$0 - 10$	$80 - 77 = 3$	5	-50	-5	-15
$10 - 20$	$77 - 72 = 5$	15	-40	-4	-20
$20 - 30$	$72 - 65 = 7$	25	-30	-3	-21
$30 - 40$	$65 - 55 = 10$	35	-20	-2	-20
$40 - 50$	$55 - 43 = 12$	45	-10	-1	-12
$50 - 60$	$43 - 28 = 15$	55	0	0	0
$60 - 70$	$28 - 16 = 12$	65	10	1	12
$70 - 80$	$16 - 10 = 6$	75	20	2	12
$80 - 90$	$10 - 8 = 2$	85	30	3	6
$90 - 100$	8	95	40	4	32
	$\sum f = 80$				$\sum fdm' = -26$

As $\overline{X} = A + \frac{\sum fdm'}{\sum f} \times C$

∴ $\overline{X} = 55 + \frac{(-26)}{80} \times 10$

$= 55 - \frac{13}{4} \quad = \frac{220 - 13}{4} \quad = \frac{207}{4} \quad = 51.75$

⇒ $\overline{X} = 51.75$

4.5.5 Arithmetic Mean in case of Unequal Class Interval Series:

Sometime the class interval between two classes is not same, for example 10-20, 20-40 etc. These series are known as unequal class interval series. However, it does not affect the finding of arithmetic mean as there is not precondition of equal class interval in case of arithmetic mean. So, mean will be calculated in usual manner.

Example 10. Calculate $\overline{X}$ if the data is given below:

C. I.	**$4 - 8$**	**$8 - 20$**	**$20 - 28$**	**$28 - 44$**	**$44 - 68$**	**$68 - 80$**
f	**3**	**8**	**12**	**21**	**10**	**6**

Solution:

C. I.	f	Mid Value m	$dm = m - A$ $A = 26$	fdm
$4 - 8$	3	6	-20	-60

$8-20$	8	14	-12	-96
$20-28$	12	24	-2	-24
$28-44$	21	36	$+10$	210
$44-68$	10	56	$+30$	300
$68-80$	6	74	$+48$	288
	$\sum f = 60$			$\sum fdm = 618$

As $\quad \overline{\times} = A + \frac{\sum fdm}{\sum f}$

⇒ $\quad \overline{\times} = 26 + \frac{618}{60} \quad = 26 + 10.3 = 36.3$

⇒ $\quad \overline{\times} = 36.3$

4.5.6 Combined Arithmetic Mean:

Sometime we have the knowledge of mean of two or more series separately but we are interested in finding the mean that will be obtained by taking all these series as one series, such mean is called combined mean. It can be calculated using the following formula.

$$\overline{\mathbf{X_{12}}} = \frac{\mathbf{N_1\overline{X_1} + N_2\overline{X_2}}}{\mathbf{N_1 + N_2}}$$

Where N_1 = Number of items in first series

N_2 = Number of of items in second series

$\overline{X_1}$ = Mean of first series

and $\quad \overline{X_2}$ = Mean of second series

Example 11. Find the combined mean for the following data

	Firm A	**Firm B**
No. of Wage Workers	**586**	**648**
Average Monthly Wage (Rs.)	**52.5**	**47.5**

Solution: Combined mean wage of all the workers in the two firms will be

$$\overline{X_{12}} = \frac{N_1\overline{X_1}+N_2\overline{X_2}}{N_1+N_2}$$

Where $\quad N_1$ = Number of workers in Firm A

N_2 = Number of workers in Firm B

$\overline{X_1}$ = Mean wage of workers in Firm A

and $\quad \overline{X_2}$ = Mean wage of workers in Firm B

We are given that

$N_1 = 586 \qquad N_2 = 648$

$\overline{X_1} = 52.5 \qquad \overline{X_2} = 47.5$

$\therefore$ Combined Mean, $\overline{X_{12}}$

$$= \frac{(586\times52.5)+(648\times47.5)}{586+648} \quad = \frac{61{,}545}{1234} \qquad = \text{Rs}.49.9$$

4.5.7 Correcting Incorrect Mean

Many a time it happens that we take some wrong items in the data or overlook some item. This results in wrong calculation of Mean. Later we find the correct values and we want to find out correct mean. This can be done using the following steps:

1. Multiply the incorrect mean of the data (incorrect $\overline{\times}$) with number of items to find out incorrect $\sum \overline{\times}$.
2. Now subtract all the wrong observation from the above values and add the correct observation to the above value to find out correct $\sum \overline{\times}$.
3. Now divide the correct $\sum \overline{\times}$. with the number of observations to find correct mean.

Example12. Mean wage of 100 workers per day found to be 75. But later on, it was found that the wages of two labourers $\text{Rs}.98$ and $\text{Rs}.69$ were misread as $\text{Rs}.89$ and $\text{Rs}.96$. Find out the correct mean wage.

Solution: We know that

$$\text{Correct} \sum X = \text{Incorrect} \sum X - (\text{Incorrect items}) + (\text{Correct Items})$$

Also $\quad \overline{X} = \frac{\sum X}{N}$

$\Rightarrow \quad \text{Incorrect} \sum X = 100 \times 75 = 7500$

$\therefore \quad \text{Correct} \sum X = 7500 - (89 + 96) + (98 + 69)$

$= 7482$

$\Rightarrow \quad \text{Correct } \overline{X} = \frac{\text{Correct } \sum X}{N}$

$= \frac{7482}{100} \quad = 74.82$

Determination of Missing Frequency

Example 13. Find the missing frequencies of the following series, if $\overline{X} = 33$ and $N = 100$

X	5	15	25	35	45	55
f	5	10	?	30	?	10

Solution: Let the missing frequencies corresponding to $X = 25$ and $X = 45$ be ‘f_1’ and ‘f_2’ respectively.

X	f	fX
5	5	25
15	10	150
25	f_1	$25f_1$
35	30	1050
45	f_2	$45f_2$
55	10	550
	$\sum f = 55 + f_1 + f_2$	$\sum fX = 1775 + 25f_1 + 45f_2$

Now, $N = 100$ (Given)

$\therefore$ $55 + f_1 + f_2 = 100$

$\Rightarrow$ $f_1 + f_2 = 45$ …(i)

Also $\overline{X} = \frac{\sum fX}{N}$

$\Rightarrow$ $33 = \frac{1775+25f_1+45f_2}{100}$

$\Rightarrow$ $3300 = 1775 + 25f_1 + 45f_2$

$\Rightarrow$ $25f_1 + 45f_2 = 1525$ …(ii)

Solving (i) and (ii), we get

$$\begin{array}{ll} 25 \times (f_1 + f_2 = 45) & \Rightarrow 25f_1 + 25f_2 = 1125 \\ 1 \times (25f_1 + 45f_2 = 1525) & \Rightarrow 25f_1 + 45f_2 = 1525 \\ & \underline{\quad (-) \quad (-) \quad (-)} \\ & \quad -20f_2 = -400 \\ & \quad f_2 = \frac{400}{20} = 20 \end{array}$$

$\therefore$ $f_2 = 20$

Put $f_2 = 20$ in (i)

$f_1 + 20 = 45$

$\Rightarrow$ $f_1 = 45 - 20 = 25$

$\therefore$ $f_1 = 25$

$\therefore$ $f_1 = 25,\ f_2 = 20$

4.5.8 Properties of Arithmetic mean

1. If we take the deviations of the observations from its Arithmetic mean and then sum up such deviations, then sum of such deviations will always be zero.
2. If we take the square of the deviations of items from its Arithmetic mean and then sum up suxh squares, the value obtained will always be less than the square of deviation taken from any other values.
3. If we have separate mean of two series, we can find the combined mean of the series.

4. If the value of all items in that data is increased or decreased by some constant value say ‘k’, then the Arithmetic mean is also increased or decreased by same ‘k’. In other words if k is added to the items then actual mean will be calculated by deducting that k from the mean calculated.
5. If value of all items in the series is divided or multiplied by some constant ‘k’ then the mean is also multiplied or divided by the same constant ‘k’. In other words, if we multiply all observations by ‘k’ then actual mean can be calculated by dividing the mean to obtained by the constant ‘k’.

4.5.9 Merits of Arithmetic Mean

1. Arithmetic mean is very simple to calculate and it is also easy to understand.
2. It is most popular method of calculating the average.
3. Arithmetic mean is rigidly defined means it has a particular formula for calculating the mean.
4. Arithmetic mean is comparatively less affected by fluctuation in the sample.
5. It is most useful average for making comparison.
6. We can perform further treatment on Arithmetic mean.
7. We need not to have grouping of items for calculating Arithmetic mean.
8. Arithmetic mean is based on all the values of the data.

4.5.10 Limitations of Arithmetic Mean

1. The biggest limitation of Arithmetic mean is that it is being affected by extreme values.
2. If we have open end series, it is difficult to measure Arithmetic mean.
3. In case of qualitative data, it is not possible to calculate Arithmetic mean.
4. Sometime it gives absurd result like we say that there are 20 students in one class and 23 students in other class then average number of students in a class is 21.5, which is not possible because student cannot be in fraction.
5. It gives more importance to large value items than small value items.
6. Mean cannot be calculated with the help of a graph.
7. It cannot be located by just inspections of the items.

CHECK YOUR PROGRESS (A)

1. Following data pertains to the monthly salaries in rupees of the employees of a Mohanta Enterprises. Calculate the average salary per employ

3000, 4100, 4700, 5400, 2300, 3400, 3700, 5100, 5300, 4700

2. Calculate mean for the following data using the shortcut method.

700, 650, 550, 750, 800, 850, 650, 700, 950

3. Following is the height of students of class tenth of a school. Find out the mean height of the students.

Height in Inches	64	65	66	67	68	69	70	71	72	73
No. of students	1	6	10	22	21	17	14	5	3	1

4. Calculate A.M for the following frequency distribution of Marks.

Marks	5	10	15	20	25	30	35	40
No of students	5	7	9	10	8	6	5	2

5. Calculate mean for the following data

Marks	5-15	15-25	25-35	35-45	45-55	55-65
No of Students	8	12	6	14	7	3

6. Calculate mean for the given data by step deviation method

C.I	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	8	12	14	16	15	9	6

7. From the following data, find the average sale per shop.

Sales in ‘000; units	10-12	13-15	16-18	19-21	22-24	25-27	28-30
No. shops	34	50	85	60	30	15	7

8. For the following data (Cumulative Series), find the average income.

Income Below in (Rs.)	30	40	50	60	70	80	90
No. of persons	16	36	61	76	87	95	100

9. Calculate the average marks for the following cumulative frequency distribution.

Marks Above	0	10	20	30	40	50	60	70	80	90
No of students	80	77	72	65	55	43	28	16	10	8

10. For a group of 50 male workers, their average monthly wage Rs.6300 and for a group of 40 female workers this average is Rs. 5400. Find the average monthly wage for the combined group of all the workers.

11. The average marks of 100 students is given to be 45. But later on, it was found that the marks of students getting 64 was misread as 46. Find the correct mean.

12. Find missing frequency when mean is 35 and number is 68.

X;	0-10	10-20	20-30	30-40	40-50	50-60
F:	4	10	12	?	20	?

13. The mean age of combined group of men and women is 30 years. The mean age of group of men is 32 years and women is 27 years. Find the percentage of men and women in the group

Answers

1) 4170	4) 20.48	7) 17.8 (in 000 units)	10) 5900	13) Men 60%
2) 733.30	5) 31.8	8) 48	11) 45.18	
3) 68.13 inches	6) 33.625	9) 51.75	12) 10,12	

4.6 MEDIAN

Median is the positional measure of Central tendency. It means the median does not depend upon the value of the item under the observation, rather it depends on the position of the item in the series. Median is a value that divide the series exactly in two equal parts, it means 50% of the observation lies below the median and 50% of the observations lies above the median. However, it is important to arrange the series either in ascending order or in descending order before calculation of Median.

For calculating Median

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.

4.6.1 Median in case of Individual series

For calculating the median in individual series, following are the steps:

1. Arrange the series in ascending or descending order.
2. Calculate the number of observations. It is denoted by N.
3. Calculate the $\left(\frac{N+1}{2}\right)^{th}$ term
4. Corresponding value to this item is the median of the data

5. In case there are even number of items in the series, this value will be in fraction. In that case take the arithmetic mean of the adjacent items in which Median is falling. For example, if it is 4.5 than take arithmetic mean of 4th item and 5th item.

$$\textbf{Median = value of } \left(\frac{\mathbf{N+1}}{\mathbf{2}}\right)^{\mathbf{th}} \textbf{ term}$$

When the number of observations N is odd

Example 1. Calculation median from the following observations:

$$\mathbf{15,\ 17,\ 19,\ 22,\ 18,\ 47,\ 25,\ 35,\ 21}$$

Solution: Arranging the given items in ascending order, we get

$$15,\ 17,\ 18,\ 19,\ 21,\ 22,\ 25,\ 35,\ 47$$

Now $\quad \text{Median, M} = \text{Size of } \left(\frac{\text{N}+1}{2}\right)^{\text{th}} \text{ item}$

$$\text{M} = \text{Size of } \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 5^{\text{th}} \text{ item}$$

$$= 21$$

$\Rightarrow \quad \text{M} = 21$

When the number of observations N is even

Example 2. Find median from the following data

$$\mathbf{28,\ 26,\ 24,\ 21,\ 23,\ 20,\ 19,\ 30}$$

Solution: Arranging the given figures in ascending order, we get

$$19,\ 20,\ 21,\ 23,\ 24,\ 26,\ 28,\ 30$$

Now $\quad \text{Median, M} = \text{Size of } \left(\frac{\text{N}+1}{2}\right)^{\text{th}} \text{ item}$

$$\text{M} = \text{Size of } \left(\frac{8+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 4.5^{\text{th}} \text{ item}$$

$$= \frac{4^{\text{th}} \text{ item} + 5^{\text{th}} \text{ item}}{2}$$

$$= \frac{23+24}{2} = \frac{47}{2} = 23.5$$

$\Rightarrow \quad \text{M} = 23.5$

4.6.2 Median in case of Discrete series

Following are the steps in case of discrete series:

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency of the series.
3. Find the $\left(\frac{\text{N}+1}{2}\right)^{\text{th}}$ term

4. Now look at this term in the cumulative frequency of the series.
5. Value against which such cumulative frequency falls is the median value.

$$\textbf{Median = value of } \left(\frac{N+1}{2}\right)^{th} \textbf{ term}$$

Example 3. Calculate the value of median, if the data is as given below:

Height (in cms.)	**110**	**125**	**250**	**200**	**150**	**180**
No. of Students	**8**	**12**	**3**	**10**	**13**	**15**

Solution: Arranging the given data in ascending order, we get

Height (in cms.)	No. of Students f	Cumulative Frequency $C \cdot f$
110	8	$8\ (1-8)$
125	12	$20\ (9-20)$
150	13	$33\ (21-33)$
180	15	$48\ (34-48)$
200	10	$58\ (49-58)$
250	3	$61\ (59-61)$
	$\sum f = N = 61$	

Now $\quad$ Median, $M = \text{Size of} \left(\frac{N+1}{2}\right)^{th}$ item

$$M = \text{Size of} \left(\frac{6+1}{2}\right)^{th} \text{ item}$$

$$= \text{Size of } 31^{st} \text{ item}$$

$$= 150$$

⇒ $\quad$ Median, $M = 150$ cms.

4.6.3 Median in case of Continuous Series

Following are the steps in case of continuous series:

1. Arrange the data in ascending or descending order.
2. Find the cumulative frequency of the series.
3. Find the $\left(\frac{N}{2}\right)^{th}$ term
4. Now look at this term in the cumulative frequency of the series. The value equal to or higher than term calculated in third step is the median class.
5. Find median using following formula.

6. $\mathbf{M = L + \frac{\frac{N}{2} - C \cdot f}{f} \times i}$

Where M = Median

L = Lower Limit of Median Class

N = Number of Observations.

c.f. = Cumulative frequency of the Median Class.

f = Frequency of the class preceding Median Class.

i = Class interval of Median Class

Example 4. Calculate Median

Marks	$\mathbf{5-10}$	$\mathbf{10-15}$	$\mathbf{15-20}$	$\mathbf{20-25}$	$\mathbf{25-30}$	$\mathbf{30-35}$
No. of Students	$\mathbf{8}$	$\mathbf{7}$	$\mathbf{14}$	$\mathbf{16}$	$\mathbf{9}$	$\mathbf{6}$

Solution:

C. I.	No. of Students f	Cumulative Frequency $C \cdot f$
$5-10$	8	$8\ (1-8)$
$10-15$	7	$15\ (9-15)$
$15-20$	14	$29\ (16-29)$
$20-25$	16	$45\ (30-45)$
$25-30$	9	$54\ (46-54)$
$30-35$	6	$60\ (55-60)$
	$\sum f = N = 60$	

$$\text{Median, } M = \text{Size of } \left(\frac{N}{2}\right)^{\text{th}} \text{ item}$$

$$M = \text{Size of } \left(\frac{60}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 30^{\text{th}} \text{ item}$$

⇒ Median lies in the class interval $20 - 25$

As Median, $M = L + \frac{\frac{N}{2} - C \cdot f}{f} \times i$

Here $L = \text{Lower limit of the median class} = 20$

$N = 60, \qquad C \cdot f = 29, \qquad f = 16$

i = Class – length of the median class = 5

∴ $M = 20 + \frac{(30-29)}{16} \times 5$

$$= 20 + \frac{5}{16}$$

$$= 20 + 9.312 = 29.312$$

$\Rightarrow$ $\qquad$ $M = 29.312$

Inclusive Series – It must be converted to Exclusive Series before calculation of the Median.

Example 5. Find Median from the given data

X	**10 − 19**	**20 − 29**	**30 − 39**	**40 − 49**	**50 − 59**	**60 − 69**	**70 − 79**	**80 − 89**
f	**6**	**53**	**85**	**56**	**21**	**16**	**4**	**4**

Solution: Converting the given data into exclusive form, we get

$$\left[\text{Correction factor} = \frac{L_2 - U_1}{2} = \frac{20 - 19}{2} = \frac{1}{2} = 0.5\right]$$

(0.5 is subtracted from all lower limits and added to all upper limits)

X	f	Cumulative frequency C · f
9.5 − 19.5	6	6 $\quad$ (1 − 6)
19.5 − 29.5	53	59 $\quad$ (7 − 59)
29.5 − 39.5	85	144 $\quad$ (60 − 144)
39.5 − 49.5	56	200 $\quad$ (145 − 200)
49.5 − 59.5	21	221 $\quad$ (201 − 221)
59.5 − 69.5	16	237 $\quad$ (222 − 237)
69.5 − 79.5	4	241 $\quad$ (238 − 241)
79.5 − 89.5	4	245 $\quad$ (242 − 245)
	$\sum f = N = 245$	

Median, $M = \text{Size of} \left(\frac{N}{2}\right)^{th}$ item

$$M = \text{Size of} \left(\frac{245}{2}\right)^{th} \text{ item}$$

$$= \text{Size of } 122.5^{th} \text{ item}$$

$\therefore$ $\qquad$ The real class limits of the median class $= (29.5 - 39.5)$

So $\qquad$ $M = L + \frac{\left(\frac{N}{2} - C\cdot f\right)}{f} \times i$

$\Rightarrow$ $\qquad$ $M = 29.5 + \left(\frac{122.5 - 59}{85}\right) \times 10$

$$= 29.5 + \left(\frac{63.5}{85} \times 10\right)$$

$$= 29.5 + \left(\frac{635}{85}\right) \qquad = 29.5 + 7.47 = 36.97$$

$\Rightarrow \qquad M = 36.97$

Cumulative Series (More than and less than)

Example 6. Find median, if the data is as given below:

Marks More than	**20**	**35**	**50**	**65**	**80**	**95**
No. of Students	**100**	**94**	**74**	**30**	**4**	**1**

Solution: Converting the given data into class – interval form, we get

Marks C. I.	Frequency f	Cumulative Frequency $C \cdot f$
$20 - 35$	$100 - 94 = 6$	$6 \quad (1 - 6)$
$35 - 50$	$94 - 74 = 20$	$26 \quad (7 - 26)$
$50 - 65$	$74 - 30 = 44$	$70 \quad (27 - 70)$
$65 - 80$	$30 - 4 = 26$	$96 \quad (71 - 96)$
$80 - 95$	$4 - 1 = 3$	$99 \quad (97 - 99)$
$95 - 110$	1	$100 \quad (100)$
	$\sum f = N = 100$	

Now $\qquad$ Median, $M = \text{Size of } \left(\frac{N}{2}\right)^{th}$ item

$$M = \text{Size of } \left(\frac{100}{2}\right)^{th} \text{ item}$$

$$= \text{Size of } 50^{th} \text{ item}$$

$\Rightarrow \qquad$ Median lies in the class interval $= 50 - 65$

So $\qquad M = L + \frac{\left(\frac{N}{2} - C \cdot f\right)}{f} \times i$

$\Rightarrow \qquad M = 50 + \left(\frac{50-26}{44}\right) \times 15$

$$= 50 + \left(\frac{24}{44} \times 15\right) \qquad = 50 + 8.18 = 58.18$$

$\Rightarrow \qquad M = 58.18$

Example 7. Find median, if the data is as given below:

Marks Less than	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**
No. of Students	**20**	**30**	**50**	**94**	**96**	**127**	**198**	**250**

Solution: Converting the given data into class interval form, we get

Marks C. I.	No. of Students f	Cumulative Frequency C · f
$0-10$	20	$20 \quad (1-20)$
$10-20$	$30-20=10$	$30 \quad (21-30)$
$20-30$	$50-30=20$	$50 \quad (31-50)$
$30-40$	$94-50=44$	$94 \quad (51-94)$
$40-50$	$96-94=2$	$96 \quad (95-96)$
$50-60$	$127-96=31$	$127 \quad (97-127)$
$60-70$	$198-127=71$	$198 \quad (128-198)$
$70-80$	$250-198=52$	$250 \quad (199-250)$
	$\sum f = N = 250$	

Now $\quad$ Median, $M = \text{Size of } \left(\frac{N}{2}\right)^{\text{th}}$ item

$$M = \text{Size of } \left(\frac{250}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 125^{\text{th}} \text{ item}$$

⇒ $\quad$ Median lies are the class – interval $= 50 - 60$

So $\quad M = L + \frac{\frac{N}{2} - C\cdot f}{f} \times i$

⇒ $\quad M = 50 + \left(\frac{125-96}{31}\right) \times 10$

$$= 50 + \left(\frac{29}{31} \times 10\right)$$

$$= 50 + \frac{290}{31} \qquad = 50 + 9.35 = 59.35$$

⇒ $\quad M = 59.35$

Mid – Value Series

Example 8. Find the value of median for the following data:

Mid Value	**15**	**25**	**35**	**45**	**55**	**65**	**75**	**85**	**95**
f	**8**	**26**	**45**	**72**	**116**	**60**	**38**	**22**	**13**

Solution: It is clear from the mid – value that the class size is 10. For finding the limits of different classes, apply the formula:

$$\mathrm{L} = \mathrm{m} - \frac{\mathrm{i}}{2} \quad \text{and} \quad \mathrm{U} = \mathrm{m} + \frac{\mathrm{i}}{2}$$

Where, L and U denote the lower and upper limits of different classes, 'm' denotes the mid – value of the corresponding class interval and 'i' denotes the difference between mid values.

∴ Corresponding to mid – value '15', we have

$$\mathrm{L} = 15 - \frac{10}{2} \quad \text{and} \quad \mathrm{U} = 15 + \frac{10}{2}$$

i.e. $\mathrm{C.I.} = 10 - 20$

Similarly other class intervals can be located

Mid Value	f	C. I.	Cumulative Frequency C · f
15	8	10 – 20	8 (1 – 8)
25	26	20 – 30	34 (9 – 34)
35	45	30 – 40	79 (35 – 79)
45	72	40 – 50	151 (80 – 151)
55	116	50 – 60	267 (152 – 267)
65	60	60 – 70	327 (268 – 327)
75	38	70 – 80	365 (328 – 365)
85	22	80 – 90	387 (366 – 387)
95	13	90 – 100	400 (388 – 400)
	N = 100		

Now Median, $\mathrm{M} = \text{Size of } \left(\frac{\mathrm{N}}{2}\right)^{\text{th}}$ item

$$\mathrm{M} = \text{Size of } \left(\frac{400}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 200^{\text{th}} \text{ item}$$

⇒ Median lies in the class – interval $= 50 - 60$

So $$\mathrm{M} = \mathrm{L} + \frac{\frac{\mathrm{N}}{2} - \mathrm{C{\cdot}f}}{\mathrm{f}} \times \mathrm{i}$$

⇒ $$\mathrm{M} = 50 + \left(\frac{200-151}{116}\right) \times 10$$

$$= 50 + \left(\frac{49}{116} \times 10\right)$$

$$= 50 + \frac{490}{116} \qquad = 50 + 4.224 = 54.224$$

⇒ $\mathrm{M} = 54.224$

Determination of Missing Frequency

Example 9. Find the missing frequency in the following distribution if $\mathbf{N = 72}$, $\mathbf{Q_1 = 25}$ and $\mathbf{Q_3 = 50}$

C. I.	**0 – 10**	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**	**60 – 70**	**70 – 80**
f	**4**	**8**	**–**	**19**	**–**	**10**	**5**	**–**

Solution: Let the missing frequencies be f_1, f_2 and f_3 respectively.

C. I.	f	Cumulative Frequency $C \cdot f$
$0 - 10$	4	4
$10 - 20$	8	12
$20 - 30$	f_1	$12 + f_1$
$30 - 40$	19	$31 + f_1$
$40 - 50$	f_2	$31 + f_1 + f_2$
$50 - 60$	10	$41 + f_1 + f_2$
$60 - 70$	5	$46 + f_1 + f_2$
$70 - 80$	f_3	$46 + f_1 + f_2 + f_3$
	$N = 72 = \sum f$ $\sum f = 46 + f_1 + f_2 + f_3$	

Now

$N = 72$

$= \sum f$

$= 46 + f_1 + f_2 + f_3$

$\Rightarrow \quad f_1 + f_2 + f_3 = 72 - 46 = 26$

$\Rightarrow \quad f_1 + f_2 + f_3 = 26$ …(i)

Also, $\quad Q_1 = 25$ (Given)

$\Rightarrow \quad Q_1$ lies in the class – interval $20 - 30$

$$\Rightarrow \quad Q_1 = L + \frac{\frac{N}{4} - C{\cdot}f}{f} \times i$$

$$25 = 20 + \frac{\frac{72}{4} - 12}{f_1} \times 10$$

$$25 = 20 + \frac{18 - 12}{f_1} \times 10$$

$$25 - 20 = \frac{6}{f_1} \times 10$$

$$5f_1 = 60$$

$$f_1 = \frac{60}{5} \quad \Rightarrow \quad f_1 = 12$$

…(ii)

Similarly, we are given that

$$Q_3 = 50$$

$\Rightarrow \quad Q_3$ lies in the class – interval $50 - 60$

$$\Rightarrow \quad Q_3 = L + \frac{\frac{3N}{4} - C{\cdot}f}{f} \times i$$

$$50 = 50 + \frac{\frac{3 \times 72}{4} - (31 + f_1 + f_2)}{10} \times 10$$

$$50 = 50 + \frac{54-(31+12+f_2)}{1} \qquad (\because f_1 = 12 \text{ By (ii)})$$

$$50 - 50 = 54 - (43 + f_2)$$

$$0 = 54 - (43 + f_2)$$

$$43 + f_2 = 54$$

$$f_2 = 54 - 43$$

$$\Rightarrow \quad f_2 = 11 \qquad \ldots\text{(iii)}$$

Putting (ii) and (iii) in (i), we get

$$f_1 + f_2 + f_3 = 26$$

$$12 + 11 + f_3 = 26$$

$$23 + f_3 = 26$$

$$f_3 = 26 - 23$$

$$\Rightarrow \quad f_3 = 3$$

4.6.4 Merits of Median

1. Median is easy to calculate.
2. It is capable of Graphic presentation.
3. It is possible even in case of open-ended series.
4. This is rigidly defined.
5. It is not affected by extreme values.
6. In case of qualitative data, it is very useful**.**

4.6.5 Limitations of Median

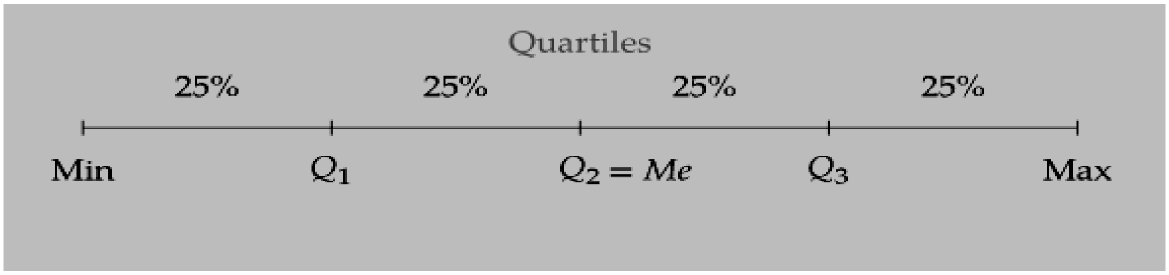
1. It is not capable of further algebraic treatment.
2. It is positional average and is not based on all observation.
3. It is very much affected by fluctuation in sampling.
4. Median needs arrangement of data before calculation.
5. In case of continuous series, it assumes that values are equally distributed in a particular class.

4.7 OTHER POSITIONAL MEASURES (QUARTILES, DECILES AND PERCENTILES)

As median divide the series into two equal parts, there are many other positional measures also. These Positional measures are also known as partition values. Following are some of the positional measure

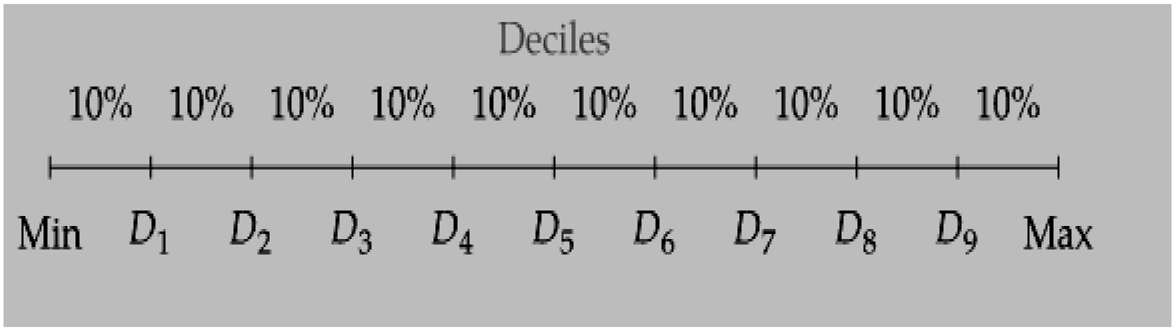
A) Quartiles

Quartile are the values that divide the series in four equal parts. There is total three quarter in number denoted by Q1, Q2 and Q3. First quartile is placed at 25% of the items, second quartile at 50% of the items, third quartile at 75% of the items. The value of second quartile is always equal to Median.



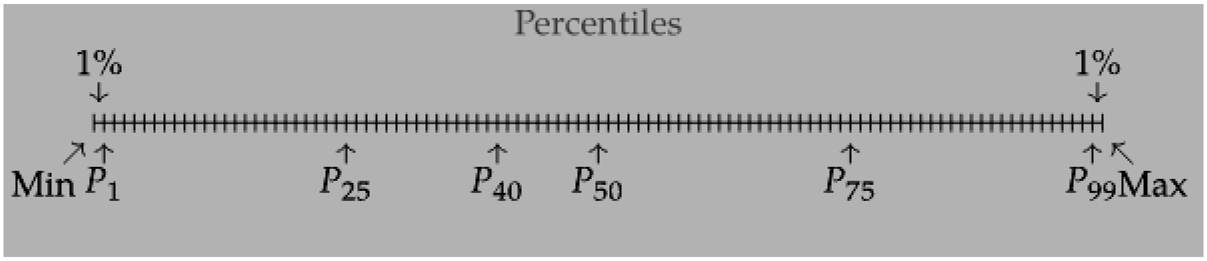
B) Deciles

Deciles are the values that divide the series in ten equal parts. There are total nine Deciles in number denoted by D1, D2, D3 and so on upto D9. The first decile is placed at 10% of the items, second quartile at 20% of the items, similarly last at 90% of the items. The value of fifth Decile is always equal to Median.



C) Percentile

Percentiles are the values that divide the series in hundred equal parts. There is total ninety-nine Percentiles in number denoted by P1, P2, P3 and so on up to P99. The first Percentile is placed at 1% of the items, second quartile at 2% of the items, similarly last at 99% of the items. The value of fiftieth Percentile is always equal to Median.



The methods of finding positional measures are same as in case of median. However, following are the formulas that can be used for finding positional measures.

CHECK YOUR PROGRESS (B)

1. Calculate Median

30, 45,75, 65, 50, 52, 28, 40, 49, 35, 52,

2. Calculate Median

36, 32,28, 22, 26, 20, 18, 40,

3. Find Median

Wages:	100	150	80	200	250	180
No. of workers:	24	26	16	20	6	30

4. Calculate Median

X:	0-5	5-10	10-15	15-20	20-25	25-30	30-35
F:	4	6	10	16	12	8	4

5. Calculate Median:

X:	10-19	20-29	30-39	40-49	50-59	60-69
F:	4	8	12	16	10	6

6. Find Median:

Income	100-200	200-400	400-700	700-1200	1200-2000
Number of firms	40	100	260	80	20

7. Find missing frequency when median is 50 and number is 100.

X:	0-20	20-40	40-60	60-80	80-100
F:	1 4	?	27	?	15

8. Find Q_1, Q_3, D_5, P_{25} and P_{67}

X: 37, 39, 45, 53, 41, 57, 43, 47, 51, 49, 55

9. Calculate Median, Quartile and D_6

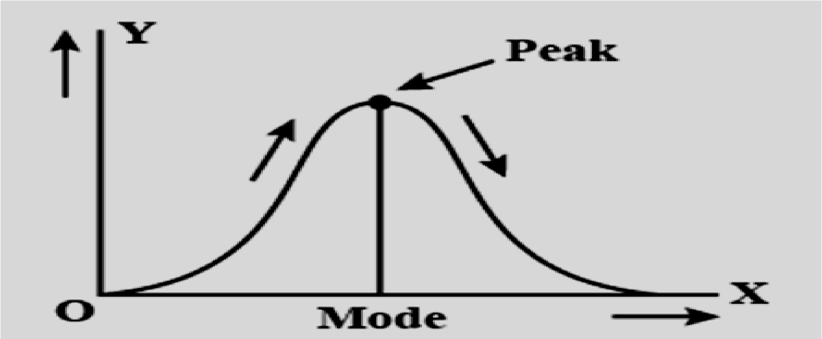
Marks Less Than	80	70	60	50	40	30	20	10
No. of Students	100	90	80	60	32	20	13	5

Answers

1) 49	2) 27	3) 150
4) 18.12	5) 42	6) 526.92
7) 23,21	8) 41, 53, 47, 41, 51.08	9) M = 46.4, Q_1 = 34.2, Q_3 = 57.5, D_6 = 50

4.8 MODE

Mode is another positional measure of Central Tendency. Mode is the value that is repeated most number of time in the series. In other words, the value having highest frequency is called Mode. The term 'Mode' is taken from French word 'La Mode' which means the most fashionable item. So, Mode is the most popular item of the series.



For calculating Mode

1. Series should be in ascending or descending order.
2. Series should be exclusive, not inclusive.
3. Series should have equal class intervals.

4.8.1 Mode in Individual Series.

In case of Individual series, following are the steps of finding the Mode.

1. Arrange the series either in ascending order or descending order.
2. Find the most repeated item.
3. This item is Mode.

Example 1. Calculate mode from the following data of marks obtained by 10 students

S. No.	1	2	3	4	5	6	7	8	9	10
Marks obtained	10	27	24	12	27	27	20	18	15	30

Solution: By Inspection

It can be observed that 27 occur most frequently i. e. 3 times. Hence, mode = 27 marks

By converting into discrete series

Marks Obtained	Frequency
10	1
12	1
15	1
18	1
20	1
24	1
27	3
30	1
	N = 10

Since, the frequency of 27 is maximum i. e. 3

It implies the item 27 occurs the maximum number of times. Hence the modal marks are 27.

$$\text{Mode} = 27$$

4.8.2 Mode in discrete series

In case of discrete series, we can find mode by two methods that are Observation Method and Grouping Method.

Observation Method: Under this method value with highest frequency is taken as mode.

Grouping Method: Following are the steps of Grouping method:

- Prepare a table and put all the values in the table in ascending order.
- Put all the frequencies in first column. Mark the highest frequency.
- In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
- In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.
- In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
- In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.
- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Value that is marked highest number of times is the mode.

Example 2. Find the modal value for the following distribution

Age (in years)	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
No. of Persons	**5**	**6**	**8**	**7**	**9**	**8**	**9**	**6**

Solution: Here, as maximum frequency 9 belongs to two age values 12 and 14, so its not possible to determine mode by inspection. We will have to determine the modal value through grouping and analysis table.

Grouping Table						
Age (in years)	Frequency					
	G_1	G_2	G_3	G_4	G_5	G_6

8	5	11	14	19	21	
9	6					24
10	8	15				
11	7		16	24		
12	9	17			26	23
13	8		17			
14	9	15				
15	6					

Analysis Table								
Group No.	8	9	10	11	12	13	14	15
G_1					×		×	
G_2					×	×		
G_3						×	×	
G_4				×	×	×		
G_5					×	×	×	
G_6			×	×	×			
Total	×	×	1	2	5	4	3	×

Since, 12 occurs maximum number of times i. e. 5 times, the modal age is 12 years

$$\text{Mode} = 12$$

4.8.3 Mode in Continuous series

In case of continuous series, we can find mode by two methods that are Observation Method and Grouping Method.

1. **Observation Method**: Under this method value with highest frequency is taken as mode class than the mode formula is applied which is given below.
2. **Grouping Method**: Following are the steps of Grouping method:
   - Prepare a table and put all the classes of data in the table in ascending order.
   - Put all the frequencies in first column. Mark the highest frequency.
   - In second column put the total of frequencies taking two frequencies at a time like first two, then next two and so on. Mark the highest total.
   - In third column put the total of frequencies taking two frequencies at a time but leaving the first frequency like second and third, third and fourth and so on. Mark the highest total.

- In fourth column put the total of frequencies taking three frequencies at a time like first three, then next three and so on. Mark the highest total.
- In fifth column put the total of frequencies taking three frequencies at a time but leaving the first frequency like second, third and fourth; than fifth, sixth and seventh and so on. Mark the highest total.
- In sixth column put the total of frequencies again taking three frequencies at a time but leaving the first two frequencies. Mark the highest total.
- Class that is marked highest number of times is the mode class.
- Apply following formula for calculating the mode:

$$\mathbf{Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i}$$

Where,

Z = Mode

L = Lower limit of the mode class

f_m = Frequency of mode class.

f_1 = Frequency of class proceeding mode class

f_2 = Frequency of class succeeding mode class

i = Class interval

Example 3. Find the mode for the following frequency distribution

Age (in years)	$\mathbf{30-35}$	$\mathbf{35-40}$	$\mathbf{40-45}$	$\mathbf{45-50}$	$\mathbf{50-55}$	$\mathbf{55-60}$
No. of Persons	**3**	**8**	**12**	**20**	**15**	**2**

Solution: Here, the maximum frequency is corresponding to the class – interval $45-50$.

So, the modal class is $45-50$.

Now, the mode is given by the formula

Mode, $Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$

Here L = Lower limit of modal class $= 45$

f_m = Frequency of modal class $= 20$

f_1 = Frequency of class preceeding the modal class $= 12$

f_2 = Frequency of class succeeding the modal class $= 15$

i = Class length of modal class $= 5$

$\therefore$ Mode, $Z = 45 + \frac{20-12}{(2\times20)-12-15} \times 5$

$= 45 + \frac{8}{40-27} \times 5$

$= 45 + 3.07$

$= 48.1$ years (approx.)

$\Rightarrow$ $Z = 48.1$ year

Example 4. Calculate mode from the following data

C. I.	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**
f	**2**	**9**	**10**	**13**	**11**	**6**	**13**	**7**	**4**	**1**

Solution: Here as it is not possible to find modal class by inspection, so we have to determine it through grouping and analysis table.

Grouping Table

C. I.	Frequency					
	G_1	G_2	G_3	G_4	G_5	G_6
$0-10$	2	11		21		
$10-20$	9		19		32	
$20-30$	10	23				34
$30-40$	13		24	30		
$40-50$	11	17			30	
$50-60$	6		19			26
$60-70$	13	20		24		
$70-80$	7		11		12	
$80-90$	4	5				
$90-100$	1					

Analysis Table

Group No.	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**	**80-90**	**90-100**
$\mathbf{G_1}$				×			×			
$\mathbf{G_2}$			×	×						
$\mathbf{G_3}$				×	×					
$\mathbf{G_4}$				×	×	×				
$\mathbf{G_5}$		×	×	×						
$\boldsymbol{G_6}$			×	×	×					
Total	×	1	3	6	3	1	1	×	×	×

Clearly the modal class is $30-40$

Now the mode is given by the formula

Mode, $Z = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$

Here $L =$ Lower limit of modal class $30 - 40 = 30$

$f_m =$ Frequency corresponding to modal class = 13

$f_1 =$ Frequency of interval preceding modal class

$f_2 =$ Frequency of interval succeeding and

$i =$ Class length of modal class

∴ Mode, $Z = 30 + \frac{13-10}{(2\times 13)-10-11} \times 10$

$= 30 + \frac{3}{26-21} \times 10$

$= 30 + \frac{30}{5}$

$= 30 + 6$

$= 36$

⇒ $Z = 36$

Example 5. Determine the missing frequencies when it is given that $N = 230$, Median, $M = 233.5$ and Mode, $Z = 234$

C.I	**200-210**	**210-220**	**220-230**	**230-240**	**240-250**	**250-260**	**260-270**
f	**4**	**16**	**–**	**–**	**–**	**6**	**4**

Solution: Let the missing frequencies be $f_1, \ f_2$ and f_3 respectively.

$C.I$	f	$C\cdot f$
$200 - 210$	4	4
$210 - 220$	16	20
$220 - 230$	f_1	$20 + f_1$
$230 - 240$	f_2	$20 + f_1 + f_2$
$240 - 250$	f_3	$20 + f_1 + f_2 + f_3$
$250 - 260$	6	$26 + f_1 + f_2 + f_3$
$260 - 270$	4	$30 + f_1 + f_2 + f_3$
	$N = 230 = \sum f$ $\sum f = 30 + f_1 + f_2 + f_3$	

Now $N = 230 = \sum f$ (Given)

$= 30 + f_1 + f_2 + f_3$

⇒ $f_1 + f_2 + f_3 = 230 - 30 = 200$

⇒ $f_1 + f_2 + f_3 = 200$ …(i)

Also, Median = 233.5 (Given)

⇒ Median class is $230 - 240$

⇒ $M = L + \frac{\frac{N}{2} - C\cdot f}{f} \times i$

$233.5 = 230 + \frac{\frac{230}{2} - (20+f_1)}{f_2} \times 10$

$3.5 = \frac{115-20-f_1}{f_2} \times 10$

$3.5f_2 = 950 - 10f_1$

⇒ $10f_1 + 3.5f_2 = 950$ …(ii)

Now Mode = 234 lies in $230 - 240$

∴ $Z = L + \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times i$

⇒ $234 = 230 + \frac{f_2 - f_1}{2f_2 - f_1 - f_3} \times 10$

⇒ $4 = \frac{f_2 - f_1}{2f_2 - f_1 - (200 - f_1 - f_2)} \times 10$ [Using (i)]

⇒ $4 = \frac{f_2 - f_1}{2f_2 - f_1 - 200 - f_1 - f_2} \times 10$

⇒ $4 = \frac{(f_2 - f_1)\times 10}{3f_2 - 200}$

⇒ $12f_2 - 800 = 10f_2 - 10f_1$

⇒ $2f_2 - 800 + 10f_1 = 0$

⇒ $10f_1 + 2f_2 = 800$ …(iii)

Solving (ii) and (iii), we get

$$\begin{array}{l} 10f_1 + 3.5f_2 = 950 \\ 10f_1 + 2f_2 \quad = 800 \\ (-) \quad (-) \quad (-) \\ \hline \qquad 1.5f_2 = 150 \end{array}$$

⇒ $f_2 = \frac{150}{1.5} = 100$

$f_2 = 100$ …(iv)

Put (iv) in (iii)

$10f_1 + 2(100) = 800$

⇒ $10f_1 = 800 - 200 = 600$

⇒ $10f_1 = 600$

⇒ $f_1 = 60$ …(v)

Put (iv) and (v) in (i)

$60 + 100 + f_3 = 200$

⇒ $f_3 = 40$

∴ The missing frequencies are 60, 100 and 40.

4.8.4 Merits of Mode

1. Mode is easy to calculate.

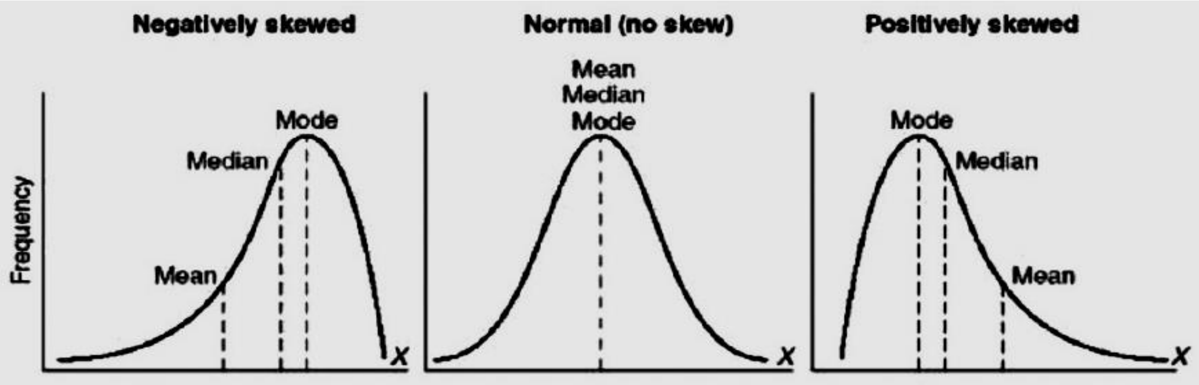
2. People can understand this in routine life.
3. It is capable of Graphic presentation.
4. It is possible even in case of open-end series.
5. This is rigidly defined.
6. It is not affected by extreme values.
7. In case of qualitative data, it is very useful**.**

4.8.5 Limitations of Median

1. It is not always determinable as series may be Bi-modal or Tri-modal.
2. It is not capable of further algebraic treatment.
3. It is positional average and is not based on all observation.
4. It is very much affected by fluctuation in sampling.
5. Mode needs arrangement of data before calculation.

4.9 RELATION BETWEEN MEAN, MEDIAN AND MODE

In a normal series the value of Mean, Median and Mode is always same. However, Karl Pearson studied the empirical relation between the Mean, Median and Mode and found that in moderately skewed series the Median always lies between the Mean and the Mode. Normally it is two third distance from Mode and one third distance from Mean.



On the basis of this relation following formula emerged

Mode = 3 Median – 2 Mean
or $\boldsymbol{Z = 3M - 2\overline{X}}$

Example 6. Calculate $\mathbf{M}$ when $\overline{\mathbf{X}}$ and Ƶ of a distribution are given to be $\mathbf{35.4}$ and $\mathbf{32.1}$ respectively.

Solution: We are given that

Mean, $\overline{\mathrm{X}} = 35.4$

Mode, $Ƶ = 32.1$

As we know the empirical relation between Mean, Median and Mode.

i.e. Mode = 3 Median – 2 Mode

$\Rightarrow \quad Ƶ = 3\mathrm{M} - 2\overline{\mathrm{X}}$

$\Rightarrow \quad \mathrm{M} = \frac{1}{3}\left(\mathrm{Z} + 2\overline{\mathrm{X}}\right)$

$\Rightarrow \quad \mathrm{M} = \frac{1}{3}\left(32.1 + 2(35.4)\right)$

$= \frac{1}{3}(32.1 + 70.8)$

$= \frac{1}{3}(102.9) = 34.3$

$\Rightarrow \quad$ Median, $\mathrm{M} = 34.3$

CHECK YOUR PROGRESS - C

1. Find Mode:

X: 22, 24, 17, 18, 19, 18, 21, 20, 21, 20, 23, 22, 22, 22

2. Find Mode by inspection method

X	6	12	18	24	30	36	42	48
f	9	11	25	16	9	10	6	3

3. Find Mode by Grouping Method

X	21	22	25	26	27	28	29	30
F	7	10	15	18	13	7	3	2

4. Find Mode by Grouping Method and inspection method

X:	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F:	2	18	30	45	35	20	6	4

5. Calculate mode using grouping and analysis methods.

X	100-110	110-120	120-130	130-140	140-150	150-160	160-170	170-180
f	4	6	20	32	33	17	8	2

6. Find Mode

X	0-100	100-200	200-400	400-500	500-700
F:	5	15	40	32	28

Answers

1. 22	2. 18	3. 26
4. 36	5. 56.46	6. 440

4.10 LET US SUM UP

- Average is the value that represent its series.
- A good average has many characteristics.
- Average is also known as Central Tendency.
- There are mainly five types of average Arithmetic Mean, Geometric Mean, Harmonic Mean, Median, Mode.
- Arithmetic mean is most popular average.
- Median divide the series in two equal parts.
- Mode is value repeated most number of time.
- There are other positional measures like Quartile, Decile and Percentile.

4.11 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Define the following:

Q1. Averages

Q2. Arithmetic mean

Q3. Formula for arithmetic mean for continues series

Q4. Combined mean formula

Q5. Median

Q6. Mode

Q7. Formula for medium for discrete series

Q8. Quartile

Q9. Percentile

Q10. Deciles

B. Long Answer Type Questions

Q1. What is central tendency. What are uses of measuring central tendency.

Q2. Give features of ideal measure of average.

Q3. What is average? Give uses and limitations of average.

Q4. What is arithmetic mean? How it is calculated.

Q5. Give properties, advantages and limitations of Arithmetic mean.

Q6. How you can calculate combined arithmetic mean?

Q7. What is median? How it is calculated in different series?

Q8. Give merits and limitations of Median.

Q9. What is mode? How it is calculated. Give its merits and limitations.

Q10. Explain grouping method of calculating Mode.

Q11. Give relation between Mean, Median and Mode.

Q12. What is Quartile, Percentile and Deciles? Explain with example.

Q13. What is positional average? Give various positional average.

Q14. According to you which measure of average is best.

4.12 FURTHER READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 5: MEASURES OF VARIATION AND SKEWNESS

STRUCTURE

5.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Explain what is Dispersion
- Compare absolute and relative measures of Dispersion
- Understand features of a good measure of Dispersion
- Calculate the Range and Quartile Deviation
- Measure the Dispersion using Mean and Standard Deviation
- Compare the variation of the two series

5.1 INTRODUCTION AND MEANING

Statistics is a tool that helps us in the extraction of information from a large pool of data. Many tools in statistics help us in extraction of data. Central tendency of data is one such tool. A good measure of central tendency is one which could represent the whole data. However, many a time we find that the average is not representing it data. The following example will make this clear:

Series X	Series Y	Series Z
100 100 100 100 100	94 105 101 98 102	1 2 3 4 490
$\sum X = 500$	$\sum Y = 500$	$\sum Z = 500$
$\bar{X} = \frac{\sum X}{N} = \frac{500}{5} = 100$	$\bar{Y} = \frac{\sum Y}{N} = \frac{500}{5} = 100$	$\bar{Z} = \frac{\sum Z}{N} = \frac{500}{5} = 100$

We can see that in all the above series the average is 100. However, in first series average is fully represents its data as all the items in the series are 100 and average is also 100. In the second series the items are very near to its average that is 100, so we can say that average is a good representation of its series. But in case of third series, average does not represent its data as there is a lot of difference between items and the average. To understand the nature of data it is very important to see the difference between items and the data. This could be done by using dispersion.

Dispersion is a very important statistical tool that help us in progress the nature of data. Dispersion shows the extent to which individual items in the data differs from its average. It is a measure of difference between data and the individual items. It indicates that how that are lacks the uniformity. Following are some of the definitions of Dispersion.

According to Simpson and Kafka, "The measures of the scatterness of a mass of figures in a series about an average is called measure of variation, or dispersion". As the dispersion gives the average difference between items and its Central tendency, it is also known as the average of second order.

5.2 BENEFITS / USES OF DISPERSION

1. **To examine reliability of Central tendency:** We have already discussed that a good measure of Central tendency is one which could represent its series. Dispersion gives us the idea that whether average is in a position to represent its series or not. Based on this, we can calculate the reliability of the average.
2. **To compare two series**: In case there are two series and we want to know that which series has more variation, we can use dispersion as its tool. In such cases normally we use relative measure of dispersion for comparing two series.
3. **Helpful in quality control**: Dispersion is tool that is frequently used in quality control by the business houses. Every manufacturer wants to maintain same quality and reduce the variation in production. Dispersion can help us in finding the deviations and removing the deviations in quality.
4. **Base of further statistical analysis:** Dispersion is a tool that is used in a number of statistical analyses. For example, we use dispersion while calculating correlation, Regression, Skewness Testing the Hypothesis etc.

5.3 FEATURES OF GOOD MEASURE OF DISPERSION

A good measure of dispersion has a number of features which are mentioned below:

1. A good tool of dispersion must be easy to understand and simple to calculate.
2. A good measure of dispersion must be based on all the values in the data.
3. It should not be affected by presence of extreme values in the data.
4. A good measure is rigidly defined.
5. A good measure of dispersion must be capable of further statistical analysis.
6. A good measure must not be affected by the sampling size.

5.4 ABSOLUTE AND RELATIVE MEASURE OF DISPERSION

Two measures of dispersion are absolute measure and relative measure:

1. **Absolute measure:** the absolute measure of dispersion is expressed in the same statistical unit in which the original values of that data are expressed. For example, if original data is represented in kilograms, the dispersion will also be represented in kilograms. Similarly, if data is represented in rupees the dispersion will also be represented in rupees. However, this measure is not useful when we have to compare two or more series that have different units of measurement or belongs to a different population.
2. **Relative measure of Dispersion**: The relative measure of dispersion is independent of unit of measurement and is expressed in pure number. Normally it is a ratio of the dispersion to the average of the data. It is very useful when we have to compare two different series that have different unit of measurement or belongs to different population.

Absolute Measure of Dispersion	**Relative Measure of Dispersion**
1. Range 2. Quartile Deviation 3. Mean Deviation 4. Standard Deviation	1. Coefficient of Range 2. Coefficient of Quartile Deviation 3. Coefficient of Mean Deviation 4. Coefficient of Standard Deviation

5.5 MEASURE OF DISPERSION - RANGE

Range is one of the simplest and oldest measures of Dispersion. We can define Range as the difference between the highest value of the data and the lowest value of the data. The more is the difference between highest and the lowest value, more is the value of Range which shows high dispersion. Similarly, less is the difference between highest and lowest value, less is value of Range which shows less dispersion. Following is formula for calculating the value of range:

$$\textbf{Range = Highest Value - Lowest Value}$$

$$\textbf{R = H – L}$$

Coefficient of Range: Coefficient of Range is relative measure of Range and can be calculated using the following formula.

$$\textbf{Coefficient of Range} = \frac{\textbf{Highest Value} - \textbf{Lowest Value}}{\textbf{Highest Value} + \textbf{Lowest Value}} = \frac{\textbf{H} - \textbf{L}}{\textbf{H} + \textbf{L}}$$

5.5.1 Range in Individual Series:

Example 1. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	330	300	470	500	410	380	425	360

Solution:

Range = Highest Value − Lowest Value

= 500 – 300 = 200

$$\text{Coefficient of Range} = \frac{\text{Highest Value} - \text{Lowest Value}}{\text{Highest Value} + \text{Lowest Value}} = \frac{500 - 300}{500 + 300} = 0.25$$

5.5.2 Range in Discrete Series:

Example 2. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	300	330	360	380	410	425	470	500
No. of Workers	5	8	12	20	18	15	13	9

Solution: Range = Highest Value − Lowest Value

= 500 – 300 = 200

$$\text{Coefficient of Range} = \frac{\text{Highest Value} - \text{Lowest Value}}{\text{Highest Value} + \text{Lowest Value}} = \frac{500 - 300}{500 + 300} = .25$$

5.5.3 Range in Continuous Series:

Example 3. Following are daily wages of workers, find out value of Range and Coefficient of Range.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution: Range = Highest Value − Lowest Value

= 90 – 10 = 80

$$\text{Coefficient of Range} = \frac{\text{Highest Value} - \text{Lowest Value}}{\text{Highest Value} + \text{Lowest Value}}$$

$$= \frac{90 - 10}{90 + 10} \qquad = .80$$

CHECK YOUR PROGRESS (A)

1. Compute for the following data Range and Coefficient of Range

28	110	27	77	19	94	63	25	111

2. Given below is the heights of students of two classes. Compare Range of the heights:

Class I:	167	162	155	180	182	175	185	158
Class II:	169	172	168	165	177	180	195	167

3. Find Range and coefficient of Range

X	5	10	15	20	25	30	35	40
f	6	4	12	7	24	21	53	47

4. Calculate coefficient of Range:

X:	10-20	20-30	30-40	40-50	50-60
F:	8	10	12	8	4

Answers

1. 92, 0.7	3. 35, 0.778
2. .088, .083	4. .714

5.6 MEASURE OF DISPERSION – QUARTILE DEVIATION

Range is simple to calculate but suffers from limitation that it takes into account only extreme values of the data and gives a vague picture of variation. Moreover, it cannot be calculated in case of open-end series. In such case we can use another method of Deviation which is Quartile Deviation or Quartile Range. Quartile Range is the difference between Third Quartile and First Quartile of the data. Following is formula for calculating Quartile Range.

$$\textbf{Quartile Range} = Q_3 - Q_1$$

Quartile Deviation: Quartile deviation is the Arithmetic mean of the difference between Third Quartile and the First Quartile of the data.

$$\textbf{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

Coefficient of Quartile Deviation: Coefficient of Quartile Deviation is relative measure of Quartile Deviation and can be calculated using the following formula.

$$\textbf{Coefficient of Range} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

5.6.1 Quartile Deviation in Individual Series:

Example 4. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	300	330	380	410	425	470	500

Solution:

Q_1 = Value of $\frac{N+1}{4}$ th item = Value of $\frac{7+1}{4}$ th item

= Value of 2n item

= 330

Q3 = Value of $\frac{3(N+1)}{4}$ th item = Value of $\frac{3(7+1)}{4}$ th item

= Value of 6th item

= 470

Quartile Range = $Q_3 - Q_1$

= 470 – 330 = 140

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

= $\frac{470 - 330}{2}$ = 70

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

= $\frac{470 - 330}{470 + 330}$ = .175

5.6.2 Quartile Deviation in Discrete Series:

Example 5. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	300	330	380	410	425	470	500
No. of Workers	5	8	12	20	18	15	13

Solution: Calculation of Quartile

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**
300	5	5
330	8	13
380	12	25
410	20	45
425	18	63

470	15	78
500	13	91

Q_1 = Value of $\frac{N+1}{4}$ th item = Value of $\frac{91+1}{4}$ th item

= Value of 23rd item

= 380

Q3 = Value of $\frac{3(N+1)}{4}$ th item = Value of $\frac{3(91+1)}{4}$ th item

= Value of 69th item = 470

Quartile Range = $Q_3 - Q_1$

= 470 – 380 = 90

Quartile Deviation = $\frac{Q_3 - Q_1}{2}$

= $\frac{470 - 380}{2}$ = 45

Coefficient of Quartile Deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1}$

= $\frac{470 - 380}{470 + 380}$ = .106

5.6.3 Quartile Deviation in Continuous Series:

Example 6. Following are daily wages of workers, find out value of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution: Calculation of Quartile

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**
10-20	5	5
20-30	8	13
30-40	12	25
40-50	20	45
50-60	18	63
60-70	15	78
70-80	13	91
80-90	9	100

Calculation of Q_1

Q_1 Class = Value of $\frac{N}{4}$ th item = Value of $\frac{100}{4}$ th item

Q_1 Class = Value of 25th item

Q_1 Class = 30-40

$$Q_1 = L_1 + \frac{\frac{n}{4} - cf}{f} \times c$$

Where $L_1 = 30$, n = 100; cf = 13; f = 12; c = 10

$$Q_1 = 30 + \frac{\frac{100}{4} - 13}{12} \times 10 \quad = 40$$

Calculation of Q_3

Q_3 Class = Value of $\frac{3N}{4}$ th item = Value of $\frac{300}{4}$ th item

Q_3 Class = Value of 75th item

Q_3 Class = 60-70

$$Q_3 = L_1 + \frac{\frac{3n}{4} - cf}{f} \times c$$

Where $L_1 = 60$, n = 100; cf = 63; f = 15; c = 10

$$Q_1 = 60 + \frac{\frac{3(100)}{4} - 63}{15} \times 10 \quad = 68$$

Calculation of Quartile Range, Quartile Deviation and Coefficient of Quartile Deviation

Quartile Range = $Q_3 - Q_1$

= 68 – 40 = 28

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{68 - 40}{2} \quad = 14$$

$$\text{Coefficient of Quartile Deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{68 - 40}{68 + 40} \quad = .259$$

5.6.4 Merits of Quartile Deviation

1. Quartile deviation is a tool that is easy to calculate and understand.
2. Quartile deviation is the best tool of dispersion in case of open-ended series.
3. This method of dispersion is better than range.
4. Unlike the range, it is not affected by the extreme values.
5. This method of dispersion is rigidly defined.
6. This method is very useful specially when we want to know the variability of middle half of the data. Under this method first 25% items that are less than Q_1 and upper 25% items that are more than Q_3 are excluded and only middle 50% items are taken.

Limitations of Quartile Deviation

1. Quartile deviation considers only middle 50% items of the data and ignore rest of the items.
2. It is not possible to make any further algebraic treatment of the quartile deviation.

3. It is not based on all the items.
4. Quartile deviation is highly affected by fluctuation in the sample.
5. It is comparatively difficult to calculate quartile deviation than range.

CHECK YOUR PROGRESS (B)

1. Find Quartile deviation and coefficient of Quartile Deviation:

X: 59, 60, 65, 64, 63, 61, 62, 56, 58, 66

2. Find Quartile deviation and coefficient of Quartile Deviation:

X	58	59	60	61	62	63	64	65	66
F	15	20	32	35	33	22	20	10	8

3. Find Quartile deviation and coefficient of Quartile Deviation

X	0-100	100-200	200-300	300-400	400-500	500-600	600-700
F:	8	16	22	30	24	12	6

4. Calculate Inter Quartile Range, Q.D and coefficient of Q.D

X	0-10	10-20	20-30	30-40	0-500	50-60	60-70	70-80	80-90
F:	11	18	25	28	30	33	22	15	22

Answers

1.	2.75, 0.0447,	2.	1.5, .024
3.	113.54, 0.335,	4.	34.84, 17.42, .3769

5.7 MEASURE OF DISPERSION – MEAN DEVIATION

Both Range and Quartile Deviation are positional method of Dispersion and takes into consideration only two values. Range considers only highest and lowest value while calculating Dispersion, while Quartile Deviation considers on First and Third Quartile for calculating Dispersion. Both these methods are not based on all the values of the data and are considerable affected by the sample unit. A good measure of Dispersion is one which considers all the values of data.

Mean Deviation is a tool of measuring the Dispersion that is based on all the values of Data. Contrary to its name, it is not necessary to calculate Mean Deviation from Mean, it can also be calculated using the Median of the data or Mode of the data. In the Mean deviation we calculated deviations of the items of data from its Average (Mean, Median or Mode) by taking positive signs only. When we divide the sum of deviation with the number of items, we get the value of Mean Deviation. In simple words:

"Mean Deviation is the value obtained by taking arithmetic mean of the deviations obtained by deducting average of data whether Mean, Median or Mode from values of data, ignoring the signs of the deviations."

5.7.1 Mean Deviation in case of Individual Series:

As we have already discussed that Mean Deviation can be calculated from Mean, Median or Mode. Following are the formula for calculating Mean Deviation in case of Individual series.

$$\textbf{Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\sum| \boldsymbol{X}-\bar{\mathbf{X}} |}{\boldsymbol{n}} = \frac{\sum| \boldsymbol{D}_{\bar{X}} |}{\boldsymbol{n}}$$

$$\textbf{Mean Deviation from Median } (MD_{M}) = \frac{\sum| \boldsymbol{X}-\boldsymbol{M} |}{\boldsymbol{n}} = \frac{\sum| \boldsymbol{D}_{\boldsymbol{M}} |}{\boldsymbol{n}}$$

$$\textbf{Mean Deviation from Mode } (MD_{Z}) = \frac{\sum| \boldsymbol{X}-\boldsymbol{Z} |}{\boldsymbol{n}} = \frac{\sum| \boldsymbol{D}_{\boldsymbol{Z}} |}{\boldsymbol{n}}$$

In case we want to calculate Coefficient of Mean Deviation, it can be done using following formulas.

$$\textbf{Coefficient of Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\mathbf{MD}_{\bar{X}}}{\bar{\mathbf{X}}}$$

$$\textbf{Coefficient of Mean Deviation from Median } (MD_{M}) = \frac{\mathbf{MD}_{M}|}{\boldsymbol{M}}$$

$$\textbf{Coefficient of Mean Deviation from Mode } (MD_{Z}) = \frac{\mathbf{MD}_{Z}}{\boldsymbol{Z}}$$

Example 7. Following are the marks obtained by Students of a class in a test. Calculated Mean Deviation from (i) Mean (ii) Median (iii) Mode. Also, calculate Coefficient of Mean Deviation.

Wage (Rs.)	5	7	8	8	9	11	13	14	15

Solution:

Let us calculate Mean Median and Mode

Mean ($\bar{X}$) = $\frac{5+7+8+8+9+11+13+14+15}{9} = \frac{90}{9} = 10$

Median (M) = Value of $\frac{N+1}{2}$ th item = Value of $\frac{9+1}{2}$ th item

= Value of 5th item = 9

Mode = Item having maximum frequency i.e., 8.

Calculation of Deviations

Marks **X**	$\mathbf{D_{\bar{X}}}$ **=│ X - $\bar{\text{X}}$│** **(Where $\bar{\text{X}}$ = 10)**	$\mathbf{D_M}$ **=│ X - M│** **(Where M = 9)**	$\mathbf{D_Z}$ **=│ X - Z│** **(Where Z = 8)**
5	5	4	3
7	3	2	1
8	2	1	0
8	2	1	0
9	1	0	1
11	1	2	3
13	3	4	5
14	4	5	6
15	5	6	7
	$\sum \mathbf{D_{\bar{X}}} = \mathbf{26}$	$\sum \mathbf{D_M} = \mathbf{25}$	$\sum \mathbf{D_Z} = \mathbf{26}$

1. Mean Deviation from Mean ($\text{MD}_{\bar{\text{X}}}$) = $\frac{\sum| \text{X}-\bar{\text{X}} |}{\text{n}} = \frac{\sum| \text{D}_{\bar{\text{X}}} |}{\text{n}} = \frac{26}{9}$ = 2.88

Coefficient of Mean Deviation from Mean ($\text{MD}_{\bar{\text{X}}}$) = $\frac{\text{MD}_{\bar{\text{X}}}}{\bar{\text{X}}} = \frac{2.88}{10}$ = .288

2. Mean Deviation from Median (MD_M) = $\frac{\sum| \text{X}-\text{M} |}{\text{n}} = \frac{\sum| \text{D}_\text{M} |}{\text{n}} = \frac{25}{9}$ = 2.78

Coefficient of Mean Deviation from Median (MD_M) = $\frac{\text{MD}_\text{M}|}{\text{M}} = \frac{2.78}{9}$ = .309

3. Mean Deviation from Mode (MD_Z) = $\frac{\sum| \text{X}-\text{Z} |}{\text{n}} = \frac{\sum| \text{D}_\text{Z} |}{\text{n}} = \frac{26}{9}$ = 2.88

Coefficient of Mean Deviation from Mode (MD_Z) = $\frac{\text{MD}_\text{Z}}{\text{Z}} = \frac{2.88}{8}$ = .36

5.7.2 Mean Deviation in case of Discrete Series:

Following are the formula for calculating Mean Deviation in case of Discrete series.

$$\boxed{\textbf{Mean Deviation from Mean (}\text{MD}_{\bar{\text{X}}}\textbf{) = } \frac{\sum \boldsymbol{f}| \boldsymbol{X}-\bar{\textbf{X}} |}{\boldsymbol{n}} = \frac{\sum \boldsymbol{f}| \boldsymbol{D}_{\bar{\text{X}}} |}{\boldsymbol{n}}}$$

$$\boxed{\textbf{Mean Deviation from Median (}\text{MD}_\text{M}\textbf{) = } \frac{\sum \boldsymbol{f}| \boldsymbol{X}-\boldsymbol{M} |}{\boldsymbol{n}} = \frac{\sum \boldsymbol{f}| \boldsymbol{D_M} |}{\boldsymbol{n}}}$$

$$\boxed{\textbf{Mean Deviation from Mode (}\text{MD}_\text{Z}\textbf{) = } \frac{\sum \boldsymbol{f}| \boldsymbol{X}-\boldsymbol{Z} |}{\boldsymbol{n}} = \frac{\sum \boldsymbol{f}| \boldsymbol{D_Z} |}{\boldsymbol{n}}}$$

Example 8. Following are the wages of workers that are employed in a factory. Calculate Mean Deviation from (i) Mean (ii) Median (iii) Mode. Also calculate Coefficient of Mean Deviation.

Wage (Rs.)	300	330	380	410	425	470	500
No. of Workers	6	8	15	25	18	15	13

Solution: Let us calculate Mean Median and Mode

X	**f**	**f X**	**cf**
300 330 380 410 425 470 500	5 8 15 26 18 15 13	1500 2640 5700 10660 7650 7050 6500	5 13 28 54 72 87 100
		$\sum X = 41700$	

Mean $(\overline{X}) = \frac{\sum X}{n} = \frac{41700}{100} = 417$

Median (M) = Value of $\frac{N+1}{2}$ th item = Value of $\frac{100+1}{2}$ th item

= Value of 50.5 item = 410

Mode = Item having maximum frequency i.e., 410.

Calculation of Deviations

X	f	$D_{\overline{X}} = \lvert X - \overline{X} \rvert$ $(\overline{X} = 417)$	$fD_{\overline{X}}$	$D_M = \lvert X - M \rvert$ $(M = 410)$	fD_M	$D_Z = \lvert X - Z \rvert$ $(Z = 410)$	fD_Z
300 330 380 410 425 470 500	5 8 15 26 18 15 13	117 87 37 7 8 53 83	585 696 555 182 144 795 1079	110 80 30 0 15 60 90	550 640 450 0 270 900 1170	110 80 30 0 15 60 90	550 640 450 0 270 900 1170
			$\sum fD_{\overline{X}} = 4036$		$\sum fD_M = 3980$	$\sum D_Z = 26$	$\sum fD_Z = 3980$

1. Mean Deviation from Mean $(MD_{\overline{X}}) = \frac{\sum f \lvert X - \overline{X} \rvert}{n} = \frac{\sum f \lvert D_{\overline{X}} \rvert}{n} = \frac{4036}{100} = 40.36$

Coefficient of Mean Deviation from Mean $(MD_{\overline{X}}) = \frac{MD_{\overline{X}}}{\overline{X}} = \frac{40.36}{417} = .097$

2. Mean Deviation from Median $(MD_M) = \frac{\sum f \lvert X - M \rvert}{n} = \frac{\sum f \lvert D_M \rvert}{n} = \frac{3980}{100} = 39.80$

Coefficient of Mean Deviation from Median $(MD_M) = \frac{MD_M \vert}{M} = \frac{39.80}{410} = .097$

3. Mean Deviation from Mode (MD_Z) = $\frac{\sum f| X- Z |}{n} = \frac{\sum f| D_Z |}{n} = \frac{3980}{100}$ = 39.80

Coefficient of Mean Deviation from Mode (MD_Z) = $\frac{MD_Z}{Z} = \frac{39.80}{410}$ = .097

5.7.3 Mean Deviation in case of Continuous Series:

In case of calculation of Mean Deviation in continuous series, the formula will remain same as we have done in Discrete Series but only difference is that instead of taking deviation from Data, we take deviations from mid value of the data. Further in case of continuous series the Mean Deviation can be calculated from Mean, Median or Mode. However, in most of cases it is calculated from Median. Following formulas are used for continuous series:

$$\textbf{Mean Deviation from Mean } (MD_{\bar{X}}) = \frac{\sum f| X- \bar{X} |}{n} = \frac{\sum f| D_{\bar{X}} |}{n}$$

$$\textbf{Mean Deviation from Median } (MD_M) = \frac{\sum f| X- M |}{n} = \frac{\sum f| D_M |}{n}$$

$$\textbf{Mean Deviation from Mode } (MD_Z) = \frac{\sum f| X- Z |}{n} = \frac{\sum f| D_Z |}{n}$$

Example 9. Following are daily wages of workers, find out value of Mean Deviation and Coefficient of Mean Deviation.

Wage (Rs.)	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Workers	5	8	12	20	18	15	13	9

Solution:

Wage (Rs.) (X)	**No. of Workers (f)**	**Cumulative Frequency (cf)**	**Mid Value (m)**	**│ D_M│ │ m – M│**	**│ f D_M│**
10-20	5	5	15	37.78	188.9
20-30	8	13	25	27.78	222.24
30-40	12	25	35	17.78	213.36
40-50	20	45	45	7.78	155.6
50-60	18	63	55	2.22	39.96
60-70	15	78	65	12.22	183.3
70-80	13	91	75	22.22	288.86
80-90	9	100	85	32.22	289.98

	N = 100				**$\sum$| fD_M| = 1582.2**

Median Class = Value of $\frac{N}{2}$ th item = Value of $\frac{100}{2}$ th item

Median Class = Value of 50th item

Median Class = 50-60

$$M = L_1 + \frac{\frac{n}{2} - cf}{f} \times c$$

Where L_1 = 50, n = 100; cf = 45; f = 18; c = 10

$$M = 50 + \frac{\frac{100}{2} - 45}{18} \times 10 \quad = 52.78$$

Calculation of Mean Deviation from Median

Mean Deviation from Median (MD_M) = $\frac{\sum f| X - M |}{n} = \frac{\sum f| D_M |}{n} = \frac{1582.2}{100}$ = 15.82

Coefficient of Mean Deviation from Median (MD_M) = $\frac{MD_M|}{M} = \frac{15.82}{52.78}$ = .30

TEST YOUR PROGRESS (C)

1. Calculate Mean Deviation from i) Mean, ii) Median, iii) Mode

X: 7, 4, 10, 9, 15, 12, 7, 9, 7

2. With Median as base calculate Mean Deviation of two series and compare variability:

Series A:	3484	4572	4124	3682	5624	4388	3680	4308
Series B:	487	508	620	382	408	266	186	218

3. Calculate Co-efficient of mean deviation from Mean, Median and Mode from the following data

X:	4	6	8	10	12	14	16
f:	2	1	3	6	4	3	1

4. Calculate Co-efficient of Mean Deviation from Median.

X;	20-25	25-30	30-40	40-45	45-50	50-55	55-60	60-70	70-80
f:	7	13	16	28	12	9	7	6	2

5. Calculate M.D. from Mean and Median

X	0-10	10-20	20-30	30-40	40-50
f	6	28	51	11	4

6. Calculate Co-efficient of Mean Deviation from Median.

X:	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60
F:	8	13	15	20	11	7	3	2	1

Answers

1) 2.35, 2.33, 2.56	3) 0.239, 0.24, 0.24	5) M.D. (Mean) 6.572, Coefficient of M.D. (Mean) 0.287, M.D. (Median) 6.4952, Coefficient of M.D. (Median) 0.281
2) 11.6%, 30.73%	4) 0.21	6) 0.22

5.8 MEASURE OF DISPERSION – STANDARD DEVIATION

Standard deviation is assumed as best method of calculating deviations. This method was given by great statistician Karl Pearson in the year 1893. In case of Mean deviation, when we take deviations from actual mean, the sum of deviations is always zero. To avoid this problem, we have to ignore the signs of the deviations. However, in case of Standard Deviation this problem is solved by taking the square of the deviations, because when we take a square of the negative sign, it is also converted into the positive sign. Then after calculating the Arithmetic mean of the deviations, we again take square root, to find out standard deviation. In other words, we can say that "Standard Deviation is the square root of the Arithmetic mean of the squares of deviation of the item from its Arithmetic mean."

The standard deviation is always calculated from the Arithmetic mean and is an absolute measure of finding the dispersion. We could also find a relative measure of standard deviation which is known as coefficient of standard deviation.

Coefficient of Standard Deviation – Coefficient of Deviation is the relative measure of the standard deviation and can be calculated by dividing the Value of Standard Deviation by the Arithmetic Mean. The value of coefficient always lies between 0 and 1, where 0 indicates no Standard Deviation and 1 indicates 100% standard deviation. Following is the formula for calculating coefficient of Standard Deviation.

$$\textbf{Coefficient of Standard Deviation} = \frac{\mathbf{SD}}{\bar{\mathbf{X}}}$$

Coefficient of Variation – Coefficient of Variation is also relative measure of the standard deviation, but unlike Coefficient of Standard Deviation it is not represented in fractions rather it is represented in terms of % age. It can be calculated by dividing the Value of Standard Deviation with the Arithmetic Mean and then multiplying resulting figure with 100. The value of coefficient always lies between 0 and 100. Following is the formula for calculating coefficient of

Standard Deviation. Low Coefficient of Variation implies less variation, more uniformity and reliability. Contrary to this higher Coefficient of Variation implies more variation, less uniformity and reliability.

$$\textbf{Coefficient of Standard Deviation} = \frac{\textbf{SD}}{\bar{\textbf{X}}} \times \textbf{100}$$

Variance – Variance is the square of the Standard Deviation. In other words, it is Arithmetic mean of square of Deviations taken from Actual Mean of the data. This term was first time used by R. A. Fischer in 1913. He used Variance in analysis of financial models. Mathematically:

$$\textbf{Variance} = (\textbf{Standard Deviation})^2 \ \textbf{ or } \sigma^2$$

5.8.1 Standard Deviation in case of Individual Series

Following are the formula for calculating Standard Deviation in case of the Individual Series:

1. **Actual Mean Method** – In this method we take deviations from actual mean of the data.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum x^2}{n}}$$

Where x = X - $\bar{X}$, n = Number of Items.

2. **Assumed Mean Method -** In this method we take deviations from assumed mean of the data. Any number can be taken as assumed mean, however for sake of simplicity it is better to take whole number as assumed mean.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum dx^2}{n} - \left(\frac{\sum dx}{n}\right)^2}$$

Where dx = X – A, n = Number of Items.

3. **Direct Methods -** In this method we don’t take deviations and standard deviation is calculated directly from the data.

$$\textbf{Standard Deviation (SD or } \sigma \textbf{)} = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

Example 10. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method. Also calculate Coefficient of Standard Deviation.

Marks	5	7	11	16	15	12	18	12

Solution:

1. Standard Deviation using Actual Mean

Marks **X**	**x = X - $\bar{X}$** **(Where $\bar{X}$ = 12)**	**x^2**
5	-7	49
7	-5	25
11	-1	01
16	4	16
15	3	09
12	0	00
18	6	36
12	0	00
$\sum$X = 96		**$\sum x^2$ = 136**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{96}{8}$ = 12

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{136}{8}} = \sqrt{17}$ =4.12

Coefficient of Standard Deviation = $\frac{SD}{\bar{X}} = \frac{4.12}{12}$ = .34

2. Standard Deviation using Assumed Mean

Marks **X**	**dx = X - A** **(Where A = 11)**	**dx^2**
5	-6	36
7	-4	16
11	0	00
16	5	25
15	4	16
12	1	01
18	7	49
12	1	01
$\sum$X = 96	**$\sum$dx = 8**	**$\sum dx^2$ = 144**

Mean ($\bar{X}$) =A+ $\frac{\sum dX}{n} = 11 + \frac{8}{8}$ = 12

Standard Deviation (**σ**) =$\sqrt{\frac{\sum dx^2}{n} - \left(\frac{\sum dx}{n}\right)^2}$ $=\sqrt{\frac{144}{8} - \left(\frac{8}{8}\right)^2} = \sqrt{18-1} = \sqrt{17}$ = 4.12

Coefficient of Standard Deviation = $\frac{SD}{\bar{X}} = \frac{4.12}{12}$ = .34

3. Standard Deviation by Direct Method

Marks (X)	**X^2**
5	25
7	49

11 16 15 12 18 12	121 256 225 144 324 144
∑X = 96	**∑X2 = 1288**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{96}{8}$ = 12

Standard Deviation (**σ**) = $\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{1288}{8} - \left(\frac{96}{8}\right)^2} = \sqrt{161 - 144} = \sqrt{17} = 4.12$

Coefficient of Standard Deviation = $\frac{SD}{\bar{X}} = \frac{4.12}{12}$ = .34

Example 11. Two Players scored following scores in 10 cricket matches. On base of their performance find out which is better scorer and also find out which player is more consistent.

Player X	26	24	28	30	35	40	25	30	45	17
Player Y	10	15	24	26	34	45	25	31	20	40

Solution: Mean and Standard Deviation of Player X

Score **X**	**x = X - $\bar{X}$** **(Where $\bar{X}$ = 30)**	**x^2**
26 24 28 30 35 40 25 30 45 17	-4 -6 -2 0 5 10 -5 0 15 -13	16 36 2 0 25 100 25 0 225 169
∑X = 300		**$\sum x^2$ = 600**

Mean ($\bar{X}$) = $\frac{\sum X}{n} = \frac{300}{10}$ = 30

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum x^2}{n}} = \sqrt{\frac{600}{10}} = \sqrt{60}$ = 7.746

Coefficient of Variation = $\frac{SD}{\bar{X}}$ × 100= $\frac{7.746}{30}$ × 100 = 25.82%

Mean and Standard Deviation of Player Y

Score (Y)	**y = Y - $\bar{Y}$ (Where $\bar{Y}$ = 27)**	**y^2**

10	-17	289
15	-12	144
24	-3	9
26	-1	1
34	7	49
45	18	324
25	-2	4
31	4	16
20	-7	49
40	13	169
$\sum X = 270$		**$\sum x^2 = 1054$**

Mean $(\overline{Y}) = \frac{\sum Y}{n} = \frac{270}{10} = 27$

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum y^2}{n}} = \sqrt{\frac{1054}{10}} = \sqrt{105.40} = 10.27$

Coefficient of Variation = $\frac{SD}{\overline{Y}} \times 100 = \frac{10.27}{27} \times 100 = 38.02\%$

Conclusion:

1. As average score of Player X is more than Player Y, he is better scorer.
2. As Coefficient of Variation of Player X is less than Player Y, he is more consistent also.

5.8.2 Standard Deviation in case of Discrete Series

Following are the formula for calculating Standard Deviation in case of the Discrete Series:

1. Actual Mean Method – In this method we take deviations from actual mean of the data.

Standard Deviation (SD or σ) = $\sqrt{\frac{\sum fx^2}{n}}$

Where x = X - $\overline{X}$, f = Frequency, n = Number of Items.

2. **Assumed Mean Method -** In this method we take deviations from assumed mean of the data.

Standard Deviation (SD or σ) = $\sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$

Where dx = X – A, n = Number of Items.

3. **Direct Methods -** In this method we don't take deviations and standard deviation is calculated directly from the data.

$$\textbf{Standard Deviation (SD or } \boldsymbol{\sigma} \textbf{) = } \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$$

Example 12. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method.

Marks	5	10	15	20	25	30	35
Frequency	2	7	11	15	10	4	1

Solution:

1. Standard Deviation using Actual Mean

Marks X	**f**	**fX**	**x = X - $\bar{X}$ ($\bar{X}$ = 19)**	**x^2**	**fx^2**
5	2	10	-14	196	392
10	7	70	-9	81	567
15	11	165	-4	16	176
20	15	300	1	1	15
25	10	250	6	36	360
30	4	120	11	121	484
35	1	35	16	256	256
	N = 50	**∑fX = 950**			**$\sum x^2$ = 2250**

Mean ($\bar{X}$) = $\frac{\sum fX}{n}$ = $\frac{950}{50}$ = 19Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum fx^2}{n}}$ = $\sqrt{\frac{2250}{50}}$ = $\sqrt{45}$ = 6.708

2. Standard Deviation using Assumed Mean

Marks X	**f**	**dx = X - A (A = 20)**	**dx^2**	**fdx**	**fdx^2**
5	2	-15	225	-30	450
10	7	-10	100	-70	700
15	11	-5	25	-55	275
20	15	0	0	0	0
25	10	5	25	50	250
30	4	10	100	40	400
35	1	15	225	15	225
	N = 50			**∑fdx = -50**	**$\sum fdx^2$ = 2300**

Standard Deviation (**σ**) $= \sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$

$= \sqrt{\frac{2300}{50} - \left(\frac{-50}{50}\right)^2}$ = $\sqrt{46 - 1}$ = $\sqrt{45}$ = 6.708

3. Standard Deviation using Direct Method

Marks (X)	**f**	$\mathbf{X^2}$	**fX**	$\mathbf{fX^2}$
5 10 15 20 25 30 35	2 7 11 15 10 4 1	25 70 225 400 625 900 1225	10 70 165 300 250 120 35	125 700 2475 6000 6250 3600 1225
	N = 50		$\sum$**fX = 950**	$\sum \mathbf{fX^2}$ **= 20300**

Standard Deviation (**σ**) $= \sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$

$= \sqrt{\frac{20300}{50} - \left(\frac{950}{50}\right)^2} = \sqrt{406 - 361} = \sqrt{45} = 6.708$

5.8.3 Standard Deviation in case of Continuous Series

In the case of continuous series, the calculation will remain same as in case of discrete series but the only difference is that instead of taking deviations from data, deviations are taken from Mid value of the data. Formulas are same as discussed above for discrete series.

Example 13. Following are the marks obtained by Students of a class in a test. Calculate Standard Deviation using (i) Actual Mean (ii) Assumed Mean (iii) Direct Method.

Also calculate coefficient of variation and Variance.

Marks	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	2	7	11	15	10	4	1

Solution: **1. Standard Deviation using Actual Mean**

Marks (X)	**m**	**f**	**fX**	**x = m -** $\mathbf{\bar{X}}$ **(**$\mathbf{\bar{X}}$ **= 21.5)**	$\mathbf{x^2}$	$\mathbf{fx^2}$
5-10 10-15 15-20 20-25 25-30 30-35 35-40	7.5 12.5 17.5 22.5 27.5 32.5 37.5	2 7 11 15 10 4 1	15 87.5 192.5 337.5 275 130 37.5	-14 -9 -4 1 6 11 16	196 81 16 1 36 121 256	392 567 176 15 360 484 256

		N = 50	**$\sum$fX = 1075**			**$\sum x^2$ = 2250**

Mean ($\bar{X}$) = $\frac{\sum fX}{n}$ = $\frac{1075}{50}$ = 21.5

Standard Deviation (SD or **σ**) = $\sqrt{\frac{\sum fx^2}{n}}$ = $\sqrt{\frac{2250}{50}}$ = $\sqrt{45}$ = 6.708

2. Standard Deviation using Assumed Mean

Marks X	**m**	**f**	**dx = X - A (A = 22.5)**	**dx^2**	**fdx**	**fdx^2**
5-10	7.5	2	-15	225	-30	450
10-15	12.5	7	-10	100	-70	700
15-20	17.5	11	-5	25	-55	275
20-25	22.5	15	0	0	0	0
25-30	27.5	10	5	25	50	250
30-35	32.5	4	10	100	40	400
35-40	37.5	1	15	225	15	225
		N = 50			**$\sum$fdx = -50**	**$\sum fdx^2$ = 2300**

Standard Deviation (**σ**) $=\sqrt{\frac{\sum fdx^2}{n} - \left(\frac{\sum fdx}{n}\right)^2}$

$=\sqrt{\frac{2300}{50} - \left(\frac{-50}{50}\right)^2} = \sqrt{46-1} = \sqrt{45} = 6.708$

3. Standard Deviation using Direct Method

Marks X	**m**	**f**	**X^2**	**fX**	**fX^2**
5-10	7.5	2	56.25	15	112.5
10-15	12.5	7	156.25	87.5	1093.75
15-20	17.5	11	306.25	192.5	3368.75
20-25	22.5	15	506.25	337.5	7593.75
25-30	27.5	10	756.25	275	7562.5
30-35	32.5	4	1056.25	130	4225
35-40	37.5	1	1406.25	37.5	1406.25
		N = 50		**$\sum$fX = 1075**	**$\sum fX^2$ = 25366.5**

Standard Deviation (**σ**) $=\sqrt{\frac{\sum fX^2}{n} - \left(\frac{\sum fX}{n}\right)^2}$

$=\sqrt{\frac{25366.5}{50} - \left(\frac{1075}{50}\right)^2} = \sqrt{507.25 - 462.25} = \sqrt{45} = 6.708$

Coefficient of Standard Deviation = $\frac{SD}{\bar{X}}$ × 100 = $\frac{6.708}{21.5}$ × 100 = 31.2%

Variance =(Standard Deviation)^2 or σ^2 = $(6.708)^2$ = 45

5.8.4 Combined Standard Deviation

The main benefit of standard deviation is that if we know the mean and standard deviation of two or more series, we can calculate combined standard deviation of all the series. This feature is not available in other measures of dispersion. That's why we assume that standard deviation is best measure of finding the dispersion. Following formula is used for this purpose:

$$\sigma_{123} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1 d_1^2 + n_2 d_2^2 + n_3 d_3^2}{n_1 + n_2 + n_3}}$$

Where,

n_1, n_2, n_3 = number of items in series 1, 2 and 3

σ_1, σ_2, σ_3 = standard deviation of series 1, 2 and 3

d_1, d_2, d_3 = difference between mean of the series and the combined mean for 1, 2 and 3.

Example14. Find the combined standard deviation for the following data

	Firm A	Firm B
No. of Wage Workers	**70**	**60**
Average Daily Wage (Rs.)	**40**	**35**
S.D of wages	**8**	**10**

Solution: Combined mean wage of all the workers in the two firms will be

$$\overline{X_{12}} = \frac{N_1\overline{X_1} + N_2\overline{X_2}}{N_1 + N_2}$$

Where N_1 = Number of workers in Firm A

N_2 = Number of workers in Firm B

$\overline{X_1}$ = Mean wage of workers in Firm A

and $\overline{X_2}$ = Mean wage of workers in Firm B

We are given that

$N_1 = 70 \qquad N_2 = 60$

$\overline{X_1} = 40 \qquad \overline{X_2} = 35$

∴ Combined Mean, $\overline{X_{12}}$

$$= \frac{(70\times40)+(60\times35)}{70+60} \qquad = \frac{4900}{130} = \text{Rs.}\,37.69$$

Combined Standard Deviation =

$$\sigma_{123} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1 d_1^2 + n_2 d_2^2}{n_1 + n_2}}$$

d_1 = 40 – 37.69 = 2.31

d_2 = 35 – 37.69 = -2.69

$$\sigma_{123} = \sqrt{\frac{70\ (8)^2 + 60\ (10)^2 + 70\ (2.31)^2 + 60\ (-2.69)^2}{70 + 60}} = 9.318$$

CHECK YOUR PROGRESS (D)

1. Calculate Standard Deviation and find Variance:

X:	5	7	11	16	15	12	18	12

2. Two Batsmen X and Y score following runs in ten matches. Find who is better Scorer and who is more consistent?

X:	26	24	28	30	35	40	25	30	45	17
Y:	10	15	24	26	34	45	25	31	20	40

3. Calculate S.D, coefficient of SD, coefficient of Variation:

X	15	25	35	45	55	65
f	2	4	8	20	12	4

4. Find Standard Deviation.

X;	5-10	10-15	15-20	20-25	25-30	30-35
F:	2	9	29	24	11	6

5. Find Standard Deviation and coefficient of variation.

X;	30-39	40-49	50-59	60-69	70-79	80-89	90-99
F:	1	4	14	20	22	12	2

6. Find Standard Deviation.

X;	0-50	50-100	100-200	200-300	300-400	400-600
F:	4	8	10	15	9	7

7. Find combined Mean and Combined Standard Deviation:

Part	No. of Items	Mean	S.D.
1	200	25	3
2	250	10	4
3	300	15	5

8. Find missing information:

	Group I	Group II	Group III	Combined
No. of Items	200	?	300	750
Mean	?	10	15	16
S.D	3	4	?	7.1924

Answers:

1) 4.12, 16.97	4) 5.74	7) 16, 7.2
2) X is better and consistent, X mean 30 CV 25.82%, Y mean 27 CV 38.02%	5) 12.505, 18.36%	8) 250, 25, 5
3) 11.83, 0.265, 26.5%	6) 141.88	

5.9 CALCULATION OF COEFFICIENT OF VARIATION

This section delves into the calculation of the CV using the formula: CV = (SD / μ) * 100, where SD is the standard deviation and μ is the mean.

The numerator and denominator components of the formula are explained in detail, ensuring a clear understanding of the calculation process. Step-by-step calculation examples are provided to illustrate the application of the CV formula. The Coefficient of Variation (CV) is calculated using a straightforward formula that involves the standard deviation (SD) and the mean (μ) of a dataset. The CV formula is as follows:

CV = (SD / μ) * 100

Where: CV: Coefficient of Variation, SD: Standard Deviation, μ: Mean

The CV formula is designed to provide a measure of relative variability by expressing the standard deviation as a percentage of the mean. It standardizes the variability metric, allowing for comparisons between datasets with different units of measurement or scales.

Let's further explain the components of the CV formula:

Standard Deviation (SD):

The standard deviation is a measure of the dispersion or variability of a dataset.

It quantifies how far individual data points deviate from the mean.

A higher standard deviation indicates greater variability in the dataset, while a lower standard deviation suggests less variability.

Mean (μ): The mean is the average value of a dataset.

It represents the central tendency or the typical value around which the data points cluster.

The mean is calculated by summing all the values in the dataset and dividing by the total number of observations.

Coefficient of Variation (CV): The CV is the ratio of the standard deviation to the mean, expressed as a percentage.

By multiplying the ratio by 100, the CV is converted into a percentage value.

The CV quantifies the relative variability of the dataset in relation to its mean, providing a standardized measure.

The CV is a dimensionless measure since it represents a ratio of two quantities with the same units. It is often expressed as a percentage to enhance its interpretability and make comparisons more intuitive.

For example, let's consider a dataset of exam scores where the mean is 80 and the standard deviation is 10. The CV can be calculated as follows:

CV = (10 / 80) * 100

CV = 12.5%

In this case, the CV of 12.5% indicates that the dataset has a relative variability of approximately 12.5% with respect to the mean. This implies that the scores exhibit moderate variability around the average performance.

Calculating the CV allows for a standardized assessment of variability, enabling comparisons between datasets with different means and standard deviations. It provides a useful measure to evaluate the relative consistency or dispersion within a dataset, supporting data analysis and decision-making processes.

To understand the Coefficient of Variation (CV) formula, it is important to grasp the meaning and significance of its numerator (standard deviation) and denominator (mean) components. Let's delve into these components in more detail:

Numerator Component: Standard Deviation (SD)

The numerator of the CV formula involves the standard deviation (SD) of the dataset. The standard deviation is a measure of the dispersion or variability of the data points from the mean. It quantifies how much individual data points deviate from the average value.

A higher standard deviation indicates greater variability in the dataset, signifying that the data points are more spread out from the mean. Conversely, a lower standard deviation suggests less variability, indicating that the data points are closer to the mean.

The standard deviation is calculated using the following steps:

Compute the mean (μ) of the dataset.

Calculate the difference between each data point and the mean.

Square each difference.

Calculate the mean of the squared differences.

Take the square root of the mean squared differences to obtain the standard deviation.

The numerator (SD) in the CV formula captures the extent of variability in the dataset, serving as a measure of dispersion.

Denominator Component: Mean (μ)

The denominator of the CV formula involves the mean (μ) of the dataset. The mean represents the average value of the dataset and serves as a measure of central tendency. It is obtained by summing all the data points and dividing the sum by the total number of observations.

The mean is a crucial component in the CV formula as it provides a reference point around which the data points are evaluated for their variability. By dividing the standard deviation by the mean, the CV expresses the variability in relation to the average value. This ratio allows for the comparison of datasets with different means and scales, making the CV a standardized measure of variability.

The CV formula, combining the standard deviation (numerator) and the mean (denominator), provides a relative measure of variability. By expressing the standard deviation as a percentage of the mean, the CV allows for meaningful comparisons across datasets.

Overall, the numerator (standard deviation) captures the dispersion or variability within the dataset, while the denominator (mean) provides a reference point for evaluating the relative variability. The combination of these components in the CV formula enables a standardized assessment of variability, facilitating comparisons and analysis across different datasets.

Example 1: Consider a dataset of monthly sales figures for a retail store over a year:

50,000, 48,000, 52,000, 55,000, 49,000, 51,000, 53,000, 50,000, 54,000, 52,000, 47,000, 50,000

Step 1: Calculate the mean (μ) of the dataset.

μ = (50,000 + 48,000 + 52,000 + 55,000 + 49,000 + 51,000 + 53,000 + 50,000 + 54,000 + 52,000 + 47,000 + 50,000) / 12

μ = 50,333.33

Step 2: Calculate the standard deviation (SD) of the dataset.

Calculate the squared difference between each data point and the mean.

Sum up the squared differences.

Divide the sum by the total number of observations (12 in this case).

Take the square root of the result.

SD = √ [((50,000 - 50,333.33) ^2 + (48,000 - 50,333.33) ^2 + ... + (50,000 - 50,333.33)^2) / 12]

SD = √ [8,666,666.67 / 12]

SD = √ [722,222.22]

SD ≈ 849.84

Step 3: Calculate the CV using the formula: CV = (SD / μ) * 100

CV = (849.84 / 50,333.33) * 100

CV ≈ 1.69%

The Coefficient of Variation (CV) for this dataset of monthly sales figures is approximately 1.69%. It indicates a relatively low level of variability in sales when compared to the mean.

Example 2: Consider a dataset of daily temperature readings in Celsius for a week:

18, 17, 16, 20, 19, 18, 17

Step 1: Calculate the mean (μ) of the dataset.

μ = (18 + 17 + 16 + 20 + 19 + 18 + 17) / 7

μ = 17.86

Step 2: Calculate the standard deviation (SD) of the dataset.

Calculate the squared difference between each data point and the mean.

Sum up the squared differences.

Divide the sum by the total number of observations (7 in this case).

Take the square root of the result.

SD = √[((18 - 17.86)^2 + (17 - 17.86)^2 + ... + (17 - 17.86)^2) / 7]

SD = √[0.48 / 7]

SD ≈ 0.30

Step 3: Calculate the CV using the formula: CV = (SD / μ) * 100

CV = (0.30 / 17.86) * 100

CV ≈ 1.68%

The Coefficient of Variation (CV) for this dataset of daily temperature readings is approximately 1.68%. It suggests a relatively low level of variability in temperature across the week when compared to the mean.

These examples illustrate the step-by-step calculation process of the Coefficient of Variation (CV) for different datasets, showcasing how the CV captures the relative variability in relation to the mean.

5.10 MEANING AND MEASURES OF SKEWNESS

Skewness is a statistical measure that helps to assess the asymmetry or lack of symmetry in a probability distribution of a random variable. It indicates the degree to which the values in a dataset are skewed or deviate from a symmetric distribution. Skewness can take positive or negative values or even zero, each indicating a different type of skewness:

a) Positive Skewness: If the distribution has a long tail on the right side and the majority of the data is concentrated on the left side, it is said to have positive skewness. The right tail is stretched out, and the mean is typically greater than the median.

b) Negative Skewness: If the distribution has a long tail on the left side and the majority of the data is concentrated on the right side, it is said to have negative skewness. The left tail is stretched out, and the mean is typically smaller than the median.

c) Zero Skewness: If the distribution is perfectly symmetrical, it has zero skewness. This means that the data is equally distributed on both sides of the mean, and the mean and median are equal.

There are various measures of skewness used to quantify the extent of skewness in a dataset. Some common measures include:

Pearson's First Coefficient of Skewness (moment skewness): It is defined as the third standardized moment of a distribution. The formula for Pearson's first coefficient of skewness is:

Skewness = (3 * (Mean - Median)) / Standard Deviation

Here, Mean refers to the arithmetic mean, Median is the median of the data, and Standard Deviation is the standard deviation of the dataset. Bowley's Skewness Coefficient: It is a measure of skewness based on quartiles. The formula for Bowley's skewness coefficient is:

Skewness = (Q1 + Q3 - 2 * Median) / (Q3 - Q1)

Here, Q1 and Q3 are the first and third quartiles, respectively.

Sample Skewness: It is a measure of skewness based on moments. The formula for sample skewness is:

Skewness = (1 / n) * ∑ [(xi - Mean) / Standard Deviation] ^3

Here, n is the sample size, xi represents each observation in the dataset, Mean is the arithmetic mean, and Standard Deviation is the standard deviation.

These are some commonly used measures of skewness, and each provides a different perspective on the skewness of the data. It's important to consider multiple measures and examine the data distribution to gain a comprehensive understanding of skewness.

5.11 SUM UP

- Dispersion shows whether average is a good representative of the series or not.
- High dispersion means values differ more than their average.
- There are two measures of dispersion, Absolute measure and relative measure.
- four methods can be used for measuring the dispersion namely, Range, Quartile Deviation, Mean Deviation and Dispersion.
- Range is a simples' method of dispersion.
- Mean deviation can be calculated from Mean, Median or Mode
- Standard Deviation is the best measure of Dispersion.
- If we know standard deviation of two series, we can calculate combined standard deviation.

5.12 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Define the terms:

Q1. Dispersion

Q2. Formula of range

Q3. Absolute measures

Q4. Name of relative measures

Q5. Quartile deviation

Q6. Standard deviation

Q7. Formula of combined standard deviation

Q8. Formula of mean deviation

B. Long Answer Type Questions

Q1. What is Dispersion? Give its uses of measuring Dispersion.

Q2. What are features of good measure of Dispersion?

Q3. What are absolute and relative measure of dispersion?

Q4. How combined standard deviation can be calculated?

Q5. Give properties of standard deviation?

5.13 FURTHER READINGS

- J. K. Sharma, Business Statistics, Pearson Education.
- S.C. Gupta, Fundamentals of Statistics, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, Elementary Statistics, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, Statistics for Management, Prentice Hall of India, New Delhi.
- M.R. Spiegel, Theory and Problems of Statistics, Schaum’s Outlines Series, McGraw Hill Publishing Co.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 6: RELATIONAL AND TREND ANALYSIS: CORRELATION AND SIMPLE REGRESSION

STRUCTURE

6.14 Suggested Readings

6.0 LEARNING OBJECTIVES

After studying the Unit, learners will be able to know about:

- Meaning of Correlation and Regression
- Different types of correlation and Regression
- Find out correlation using the graphic method
- Calculate correlation by Karl Pearson Method
- Measure correlation using the rank correlation method
- Lines of regression from x to y and y to x
- Properties of correlation and regression

6.1 INTRODUCTION: MEANING OF CORRELATION

Correlation is a statistical technique that studies the relation between two or more variables. It studies that how to variables are related to each other. It studies how change in value of one variable affects the other variable, for example in our daily life we will find the relation between income and expenditure, income and demand, Price and Demand age of husband-and-wife etcetera correlation helps in understanding such relations of different variables two variables are said to be related to each other when change in value of one variable so results in to change in value of other variable.

According to W.I. King, “Correlation means that between two series or groups of data there exists some casual connection.”

6.2 TYPES OF CORRELATION

1. **Positive correlation**: It is a situation in which two variables move in the same direction. In this case if the value of one variable increases the value of other variable also increase. Similarly, if the value of one variable decrease, the value of other variable also decrease. So, when both the variables either increase or decrease, it is known as positive correlation. For example, we can find Positive correlation between Income and Expenditure, Population and Demand of food products, Incomes and Savings etc. Following data shows positive

correlation between two variables:

Height of Persons: X	158	161	164	166	169	172	174
Weight of Person: Y	61	63	64	66	67	69	72

2. **Negative or Inverse Correlation**: When two variables move in opposite direction from each other, it is known as negative or inverse correlation. In other words, we can say that when the value of one variable increase value of other variable decrease, it is called negative correlation. In our life we find negative correlation between a number of variables, for example there is negative correlation between Price and Demand, Number of Workers and Time required to complete the work etc. Following data shows the negative correlation between two variables:

Price of Product: X	1	2	3	4	5
Demand of Product: Y	50	45	40	35	30

3. **Zero or No Correlation:** When two variables does not show any relation, it is known as zero or no correlation. In other words, we can say that in case of zero correlation, the change in value of one variable does not affect the value of another variable. In this case two variables are independent from each other. For example, there is zero correlation between height of the student and marks obtained by the student.

4. **Simple Correlation**: When we study relation between two variables only, it is known as simple correlation. For example, relation between income and expenditure, Price and Demand, are situations of simple correlation.

5. **Multiple Correlation**: Multiple correlation is a situation in which more than two variables are involved. Here relation between more than two variables is studied together, for example if we are studying the relation between income of the consumer, price if the product and demand of the product, it is a situation of multiple correlation. In case we study relation of more than two variables and all the variables are taken together, it is a situation of total correlation. For example, if we are studying the relation between the income of the consumer, price of the product and demand of the product, taking all the factors together it is called total correlation.

6. **Partial Correlation**: In case of partial correlation more than two variables are involved, but while studying the correlation we take only two factors in consideration assuming that the value of other factors is constant. For example, while studying the relation between income of the consumer, price of the product and demand of the product, we take into consideration only relation between price of the product and demand of the product assuming that income of the

consumer is constant.

7. Linear Correlation: When the change in value of one variable results into constant ratio of change in the value of other variable, it is called linear correlation. In such case if we draw the values of two variables on the graph paper, all the points on the graph paper will fall on a straight line. For example, every change in income of consumer by Rs. 1000 results into increase in consumption by 10 kg., is known as linear correlation. Following data shows example of linear correlation:

Price of Product: X	1	2	3	4	5
Demand of Product: Y	50	45	40	35	30

8. Non - Linear Correlation: When the change in value of one variable does not result into constant ratio of change in the value of other variable, it is called nonlinear correlation. In such case, if we draw the value of two variables on the graph paper all the points will not fall in the straight line on the graph. Following data shows nonlinear correlation between two variables:

Price of Product: X	1	2	3	4	5
Demand of Product: Y	50	40	35	32	30

6.3 PROPERTIES OF CORRELATION

1. Range: The coefficient of Correlation always lies between -1 to +1.
2. Degree Of Measurement: Correlation Coefficient is independent of units of measurement.
3. Direction: The sign of Correlation is positive (+ve) if the values of variables move in the same direction, if -ve then the opposite direction.
4. Symmetry: Correlation Coefficient deals with the property of symmetry. It means $r_{xy}=r_{yx}$,
5. Geometric Mean: The coefficient of Correlation is also the geometric mean of two regression coefficients Rxy= bxy. byx
6. If x and y are independent then $r_{xy}=0$
7. Change of Origin: The correlation coefficient is independent of change of origin
8. Change of Scale: The correlation coefficient is independent of change in Scale
9. Coefficient of determination: The square of the correlation coefficient (r_{xy}) is known as the coefficient of determination

6.4 KARL PEARSONS'S COEFFICIENT OF CORRELATION

Karl Peason's Coefficient of Correlation is the most important method of measuring the correlation. He was the first person who introduced the mathematical model of finding the correlation. Karl Peason's Coefficient of correlation is also denoted as 'Product Moment Correlation' also. The coefficient of correlation given by Karl Pearson is denoted as a symbol 'r'. It is the relative measure of finding the correlation. According to Karl Pearson we can determine correlation by dividing the product of deviations taken from mean of the data.

6.4.1 Direct method of calculating Correlation

Correlation can be calculated using the direct method without taking any mean. Following are the steps:

1. Take two series X and Y.
2. Find the sum of these two series denoted as ∑X and ∑Y.
3. Take the square of all the values of the series X and series Y.
4. Find the sum of the square so calculated denoted by $\sum X^2$ and $\sum Y^2$.
5. Multiply the corresponding values of series X and Y and find the product.
6. Sum up the product so calculated denoted by ∑ X Y.
7. Apply the following formula for calculating the correlation.

Coefficient of Correlation, $$\boldsymbol{r = \frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}}$$

Example 1. ***Calculate the Karl Pearson's coefficient of correlation for the following data***

X	2	3	1	5	6	4
Y	4	5	3	4	6	2

Solution:

X	Y	X^2	Y^2	XY
2	4	4	16	8
3	5	9	25	15
1	3	1	9	3
5	4	25	16	20
6	6	36	36	36
4	2	16	4	8
$\sum X = 21$	$\sum Y = 24$	$\sum X^2 = 91$	$\sum Y^2 = 106$	$\sum XY = 90$

$N = 6$

Coefficient of Correlation, $r = \frac{N\sum XY-(\sum X)(\sum Y)}{\sqrt{N\sum X^2-(\sum X)^2}\sqrt{N\sum Y^2-(\sum Y)^2}}$

$$= \frac{6\times 90-21\times 24}{\sqrt{6\times 91-(21)^2}\sqrt{6\times 106-(24)^2}}$$

$$= \frac{540-504}{\sqrt{546-441}\sqrt{636-576}}$$

$$= \frac{36}{\sqrt{105}\sqrt{60}} \qquad = \frac{36}{10.246\times 7.7459} \qquad = \frac{36}{79.31} = 0.4539$$

$\Rightarrow \qquad r = 0.4539$

6.4.2 Actual Mean method of calculating Correlation

Under this Correlation is calculated by taking the deviations from actual mean of the data. Following are the steps:

1. Take two series X and Y.
2. Find the mean of both the series X and Y, denoted by $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$.
3. Take deviations of series X from it mean and it is denoted by ‘x’.
4. Take deviations of series Y from it mean and it is denoted by ‘y’.
5. Take square of deviation of series X denoted by x^2.
6. Sum up square of deviations of series X denoted by $\sum x^2$.
7. Take square of deviation of series Y denoted by y^2.
8. Sum up square of deviations of series Y denoted by $\sum y^2$.
9. Find the product of x and y and it is denoted by xy.
10. Find the sum of ‘xy ‘it is denoted by $\sum$ xy
11. Apply the following formula for calculating the correlation.

$$\boldsymbol{r} = \frac{\sum \boldsymbol{xy}}{\sqrt{\sum \boldsymbol{x^2}}\sqrt{\sum \boldsymbol{y^2}}} = \frac{\sum(\boldsymbol{X}-\bar{\mathbf{X}})(\boldsymbol{Y}-\bar{\mathbf{Y}})}{\sqrt{\sum(\boldsymbol{X}-\bar{\mathbf{X}})^2}\sqrt{(\boldsymbol{Y}-\bar{\mathbf{Y}})^2}}$$

Example 2. ***Calculate Karl Pearson’s coefficient of correlation***

$\boldsymbol{X}$	**50**	**50**	**55**	**60**	**65**	**65**	**65**	**60**	**60**	**50**
$\boldsymbol{Y}$	**11**	**13**	**14**	**16**	**16**	**15**	**15**	**14**	**13**	**13**

Solution: When deviations are taken from actual arithmetic mean, ‘r’ is given by

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

Where $x = X - \bar{\mathrm{X}} =$ Deviation from $A.M.$ of X series

$y = Y - \bar{\mathrm{Y}} =$ Deviation from $A.M.$ of Y series

X	Y	$x = (X - \overline{X})$	x^2	$y = (Y - \overline{Y})$	y^2	xy
50	11	-8	64	-3	9	24
50	13	-8	64	-1	1	8
55	14	-3	9	0	0	0
60	16	2	4	2	4	4
65	16	7	49	2	4	14
65	15	7	49	1	1	7
65	15	7	49	1	1	7
60	14	2	4	0	0	0
60	13	2	4	-1	1	-2
50	13	-8	64	-1	1	8
$\sum X = 580$	$\sum Y = 140$		$\sum x^2 = 360$		$\sum y^2 = 22$	$\sum xy = 70$

Here, $N = 10$

$$A.M. \text{ of } X \text{ series}, \overline{X} = \frac{\sum X}{N} = \frac{580}{10} = 58$$

$$A.M. \text{ of } Y \text{ series}, \overline{Y} = \frac{\sum Y}{N} = \frac{140}{10} = 14$$

Coefficient of Correlation, $r = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{70}{\sqrt{360 \times 22}} = \frac{70}{\sqrt{7920}} = 0.7866$

$\Rightarrow$ $r = 0.7866$

6.4.3 Assumed Mean method of calculating Correlation

Under this Correlation is calculated by taking the deviations from assumed mean of the data. Following are the steps:

1. Take two series X and Y.
2. Take any value as assumed mean for series X.
3. Take deviations of series X from its assumed mean and it is denoted by ‘dx’.
4. Find sum of deviations denoted by ∑dx.
5. Take square of deviation of series X denoted by dx^2
6. Sum up square of deviations of series X denoted by ∑ dx^2.
7. Take any value as assumed mean for series Y .
8. Take deviations of series Y from its assumed mean and it is denoted by ‘dy’.
9. Find sum of deviations of series Y denoted by ∑dy.

10. Take square of deviation of series Y denoted by dy^2

11. Sum up square of deviations of series Y denoted by ∑ dy^2.

12. Find the product of dx and dy and it is denoted by dxdy.

13. Find the sum of ‘dxdy ‘it is denoted by ∑ dxdy

14. Apply the following formula for calculating the correlation.

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

Example 3. Compute coefficient of correlation from the following figures

City	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Population (in thousands)	**78**	**25**	**16**	**14**	**38**	**61**	**30**
Accident Rate (per million)	**80**	**62**	**53**	**60**	**62**	**69**	**67**

Solution: Here, $N = 7$

Coefficient of Correlation, r is given by

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

Where $dx =$ Deviations of terms of X series from assumed mean $A_X = X - A_X$

$dy =$ Deviations of terms of Y series from assumed mean $A_Y = Y - A_Y$

X	Y	dx $= X - A_X$ $A_X = 38$	$dy = Y - A_Y$ $A_Y = 67$	dx^2	dy^2	$dxdy$
70	80	32	13	1024	169	416
25	62	-13	-5	169	25	65
16	53	-22	-14	482	196	308
14	60	-24	-7	576	49	168
38	62	0	-5	0	25	0
61	69	23	2	529	4	46
30	67	-8	0	64	0	0
		$\sum dx$ $= -12$	$\sum dy$ $= -16$	$\sum dx^2$ $= 2846$	$\sum dy^2$ $= 468$	$\sum dxdy$ $= 1003$

Here, $N = 7$

∴ Coefficient of Correlation, $r = \frac{7\times1003-(-12)(-16)}{\sqrt{7\times2846-(-12)^2}\,\sqrt{7\times468-(-16)^2}}$

$$= \frac{7021-192}{\sqrt{19{,}922-144}\,\sqrt{3276-256}}$$

$$= \frac{6829}{\sqrt{19{,}778}\,\sqrt{3020}} = 0.8837$$

$\Rightarrow$ $$r = 0.8837$$

6.4.4 Step Deviation method of calculating Correlation

Under this method assumed mean is taken but the difference is that after taking the deviation, these are divided by some common factor to get the step deviations. Following are the steps:

1. Take two series X and Y.
2. Take any value as assumed mean for series X.
3. Take deviations of series X from its assumed mean and it is denoted by ‘dx’.
4. Divide the value of ‘dx’ so obtained by some common factor to get $\boldsymbol{dx'}$
5. Find sum of deviations denoted by $\sum\boldsymbol{dx'}$.
6. Take square of deviation of series X denoted by $\boldsymbol{dx'}^{\,2}$
7. Sum up square of deviations of series X denoted by $\sum \boldsymbol{dx'}^{\,2}$.
8. Take any value as assumed mean for series Y.
9. Take deviations of series Y from its assumed mean and it is denoted by ‘dy’.
10. Divide the value of ‘dy’ so obtained by some common factor to get $\boldsymbol{dy'}$
11. Find sum of deviations of series Y denoted by $\sum\boldsymbol{dy'}$.
12. Take square of deviation of series Y denoted by $\boldsymbol{dy'}^{\,2}$
13. Sum up square of deviations of series Y denoted by $\sum \boldsymbol{dy'}^{\,2}$.
14. Find the product $\boldsymbol{dx'}$ of and $\boldsymbol{dy'}$ and it is denoted by $\boldsymbol{dx'}\ \boldsymbol{dy'}$.
15. Find the sum of ‘dxdy ‘it is denoted by $\sum \boldsymbol{dx'}\ \boldsymbol{dy'}$
16. Apply the following formula for calculating the correlation.

Coefficient of Correlation, $$\boldsymbol{r} = \frac{\boldsymbol{N\sum dx'dy' - (\sum dx')(\sum dy')}}{\sqrt{\boldsymbol{N\sum dx'^2 - (\sum dx')^2}}\ \sqrt{\boldsymbol{N\sum dy'^2 - (\sum dy')^2}}}$$

Example 4. Find the coefficient of correlation by Karl Pearson’s method

Price $(\boldsymbol{Rs.})$	**5**	**10**	**15**	**20**	**25**
Demand $(\boldsymbol{kg})$	**40**	**35**	**30**	**25**	**20**

Solution:

X	Y	$dx =$ $X - A$ $A = 15$	$dx' = \frac{dx}{C_1}$ $C_1 = 5$	$dy =$ $Y - B$ $B = 30$	$dy' = \frac{dy}{C_1}$ $C_2 = 5$	dx'^2	dy'^2	$dx'dy'$
5	40	-10	-2	10	2	4	4	-4

10	35	-5	-1	5	1	1	1	-1
15	30	0	0	0	0	0	0	0
20	25	5	1	-5	-1	1	1	-1
25	20	10	2	-10	-2	4	4	-4
			$\sum dx' = 0$		$\sum dy' = 0$	$\sum dx'^2 = 10$	$\sum dy'^2 = 10$	$\sum dx'dy' = -10$

Here $N = 5$

Coefficient of Correlation, $r = \frac{N\sum dx'dy' - (\sum dx')(\sum dy')}{\sqrt{N\sum dx'^2 - (\sum dx')^2}\sqrt{N\sum dy'^2 - (\sum dy')^2}}$

$$= \frac{5\times(-10)-0\times0}{\sqrt{5\times10-0^2}\,\sqrt{5\times10-0^2}}$$

$$= \frac{-50}{\sqrt{50}\times\sqrt{50}} = -1$$

$\Rightarrow$ $r = -1$

6.4.5 Calculating Correlation with help of Standard Deviations

Under this method assumed mean is taken but the difference is that after taking the deviation, these are divided by some common factor to get the step deviations. Following are the steps:

1. Take two series X and Y.
2. Find the mean of both the series X and Y, denoted by $\bar{\mathbf{X}}$ and $\bar{\mathbf{Y}}$.
3. Take deviations of series X from it mean and it is denoted by ‘x’.
4. Take deviations of series Y from it mean and it is denoted by ‘y’.
5. Find the product of x and y and it is denoted by xy.
6. Find the sum of ‘xy ‘it is denoted by $\sum$ xy
7. Calculate the standard deviation of both series X and Y.
8. Apply the following formula for calculating the correlation.

$$\boldsymbol{r = \frac{\sum xy}{N\sigma_X\sigma_Y}}$$

Example 5. ***Given***

No. of pairs of observations $\boldsymbol{= 10}$

$\boldsymbol{\sum xy = 625}$

X Series Standard Deviation $\boldsymbol{= 9}$

Y Series Standard Deviation =8

Find ‘r’.

Solution: We are given that

$N = 10,$ $\sigma_X = 9$ $\sigma_Y = 8$ and $\sum xy = 625$

Now $r = \frac{\sum xy}{N\sigma_X\sigma_Y}$

$$= \frac{625}{10\times9\times8} \quad = \frac{625}{720} \quad = 0.868$$

$\Rightarrow$ $r = +.868$

Example 6. ***Given***

No. of pairs of observations = $\mathbf{10}$

X Series Arithmetic Mean = $\mathbf{75}$

Y Series Arithmetic Mean = $\mathbf{125}$

X Series Assumed Mean = $\mathbf{69}$

Y Series Assumed Mean = $\mathbf{110}$

X Series Standard Deviation = $\mathbf{13.07}$

Y Series Standard Deviation = $\mathbf{15.85}$

Summation of products of corresponding deviation of X and Y series = $\mathbf{2176}$

Find 'r'.

Solution: We are given that

$$N = 10, \quad \overline{X} = 75, \quad A_X = 69, \quad \sigma_X = 13.07$$

$$\overline{Y} = 125, \quad A_Y = 110, \quad \sigma_Y = 15.85$$

and $\sum xy = 2176$

Now $r = \frac{\sum xy - N(\overline{X}-A_X)(\overline{Y}-A_Y)}{N\sigma_X\sigma_Y}$

$$= \frac{2176-10(75-69)(125-110)}{10\times13.07\times15.85} \quad = \frac{2176-900}{2071.595}$$

$$= 0.6159 \approx 0.616$$

$\Rightarrow$ $r = +0.616$

Example 7. A computer while calculating the coefficient of correlation between the variables X and Y obtained the values as

$$\mathbf{N = 6, \quad \sum X = 50, \quad \sum X^2 = 448}$$

$$\mathbf{\sum Y = 106, \quad \sum Y^2 = 1896, \quad \sum XY = 879}$$

But later on, it was found that the computer had copied down two pairs of observations as

X	Y
5	15
10	18

While the correct values were

X	Y
6	18
10	19

Find the correct value of correlation coefficient.

Solution: Incorrect value of $\sum X = 50$

∴ Correct value of $\sum X = 50 - 5 - 10 + 6 + 10$

$= 51$

Incorrect value of $\sum Y = 106$

∴ Correct value of $\sum Y = 106 - 15 - 18 + 18 + 19$

$= 110$

Incorrect value of $\sum X^2 = 448$

∴ Correct value of $\sum X^2 = 448 - 5^2 - (10)^2 + 6^2 + (10)^2$

$= 459$

Incorrect value of $\sum Y^2 = 1896$

∴ Correct value of $\sum Y^2 = 1896 - 15^2 - (18)^2 + (18)^2 + 19^2$

$= 2032$

Incorrect value of $\sum XY = 879$

∴ Correct value of $\sum XY = 879 - (5 \times 15) - (10 \times 18) + (6 \times 18) + (10 \times 19)$

$= 952$

Thus, the corrected value of coefficient of correlation

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

$$= \frac{6\times952 - 51\times110}{\sqrt{6\times459 - (51)^2}\,\sqrt{6\times2032 - (110)^2}}$$

$$= \frac{5712 - 5610}{\sqrt{2754 - 2601}\,\sqrt{12{,}192 - 12{,}100}} \quad = \frac{102}{\sqrt{153}\,\sqrt{92}}$$

$$= \frac{102}{12.369\times9.59} \quad = \frac{102}{118.618} = 0.8599$$

⇒ $r = +0.8599$

TEST YOUR UNDERSTANDING (A)

1. From the following data of prices of product X and Y draw scatter diagram.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Price of X	60	65	65	70	75	75	80	85	80	100
Price of Y	120	125	120	110	105	100	100	90	80	60

2. Calculate Karl Pearson's coefficient of correlation

X	21	22	23	24	25	26	27	28	29	30
Y	46	42	38	34	30	26	22	18	14	10

3**.** Calculate Karl Pearson's coefficient between X and Y

X	42	44	58	55	89	98	66
Y	56	49	53	58	65	76	58

4**.** Find correlation between marks of subject A Subject B

Subject A	24	26	32	33	35	30
Subject B	15	20	22	24	27	24

5. Find correlation between Height of Mother and Daughter

Height of Mother (Inches)	54	56	56	58	62	64	64	66	70	70
Height of Daughter (Inches)	46	50	52	50	52	54	56	58	60	62

6. What is the Karl Pearson's coefficient of correlation if $\sum\ x\,y\ = 40\ , n =\ 100\,, \sum x^2\ = 80\ and\ \sum y^2\ = 20.$

7. Calculate the number of items for which r = 0.8, $\sum xy\ = 200$. Standard deviation of y = 5 and $\sum x^2\ = 100$ where x and y denote the deviations of items from actual means.

8. Following values were obtained during calculation of correlation:

N = 30; $\sum X = 120$ $\sum X^2 = 600$ $\sum Y = 90$ $\sum Y^2 = 250$ $\sum XY = 335$

Later found that two pairs were taken wrong which are as follows:

pairs of observations as:	(X, Y):	(8, 10)	(12, 7)
While the correct values were:	(X, Y):	(8, 12)	(10, 8)

Find correct correlation.

9. From the data given below calculate coefficient of correlation.

	X series	Y series
Number of items	8	8
Mean	68	69
Sum of squares of deviation from mean	36	44
Sum, of product of deviations x and y from means	24	24

10. Find the correlation between age and playing habits from the following data:

Age	15	16	17	18	19	20
No of students	20	270	340	360	400	300
Regular players	150	162	170	180	180	120

Answers

	5. .95	9. .603
2. -1	6. 1	10. -.94
3. .9042	7. 25	
4. .92	8. -.4311	

6.5 SPEARMAN'S RANK CORRELATION

Karl Peason's Coefficient of Correlation is very useful if data is quantitative, but in case of qualitative data it is a failure. Spearman's Rank correlation is a method that can calculate correlation both from quantitative and qualitative data if the data is ranked like in singing contest we rank the participants as one number, two number or three number etc. This method was given by Charles Edward Spearman in 1904. In this method we give Rank to the data and with help of such ranks, correlation is calculated.

6.5.1 Spearman's Rank Correlation when ranks are given.

1. Calculate the difference between ranks of both the series denoted by ∑D.
2. Take square of deviations and calculate the value of D^2.
3. Calculate sum of square of deviations denoted by $\sum D^2$.
4. Apply following formula.

Example 9. ***Following are given the ranks of* 8 *pairs. Find 'r'***

Rank X	**6**	**4**	**8**	**2**	**7**	**5**	**3**	**1**
Rank Y	**4**	**8**	**7**	**3**	**6**	**5**	**1**	**2**

Solution:

Rank X	Rank Y	Difference of Ranks D	D^2
6	4	$+2$	4
4	8	-4	16
8	7	-1	1
2	3	-1	1
7	6	$+1$	1
5	5	0	0
3	1	$+2$	4
1	2	-1	1
$N = 8$			$\sum D^2 = 28$

Coefficient of Rank Correlation, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$$= 1 - \frac{6\times 28}{8(8^2-1)}$$

$$= 1 - \frac{168}{8(64-1)}$$

$$= 1 - \frac{168}{8(63)}$$

$$= 1 - \frac{168}{504} = 1 - 0.33 = 0.67$$

⇒ Rank Correlation Coefficient $= 0.67$

Example 10. ***In a beauty contest, three judges gave the following ranks to* 10 *contestants. Find out which pair of judges agree or disagree the most.***

***Judge* 1**	**5**	**1**	**6**	**3**	**8**	**7**	**10**	**9**	**2**	**4**
***Judge* 2**	**9**	**7**	**10**	**5**	**8**	**4**	**3**	**6**	**1**	**2**

***Judge* 3**	**6**	**4**	**7**	**10**	**5**	**3**	**1**	**9**	**2**	**8**

Solution :

Ranks by			$D_1 = R_1 - R_2$	${D_1}^2$	$D_2 = R_2 - R_3$	${D_2}^2$	$D_3 = R_1 - R_3$	${D_3}^2$
Judge 1 R_1	Judge 2 R_2	Judge 3 R_3						
5	9	6	-4	16	3	9	-1	1
1	7	4	-6	36	3	9	-3	9
6	10	7	-4	16	3	9	-1	1
3	5	10	-2	4	-5	25	-7	49
8	8	5	0	0	3	9	3	9
7	4	3	$+3$	9	1	1	4	16
10	3	1	$+7$	49	2	4	9	81
9	6	9	$+3$	9	-3	9	0	0
2	1	2	$+1$	1	-1	1	0	0
4	2	8	$+2$	4	-6	36	-4	16
				$\sum {D_1}^2 = 144$		$\sum {D_2}^2 = 112$		$\sum {D_3}^2 = 182$

Now $r_{12} = 1 - \frac{6\sum {D_1}^2}{N(N^2-1)}$

$= 1 - \frac{6\times 144}{10(10^2-1)}$

$= 1 - \frac{864}{10(100-1)} \quad = 1 - \frac{864}{10(99)} \quad = 1 - \frac{864}{990}$

$= 1 - 0.873 = 0.127$

∴ $r_{12} = +0.127 \Rightarrow$ Low degree $+ve$ correlation

$r_{23} = 1 - \frac{6\sum {D_2}^2}{N(N^2-1)}$

$= 1 - \frac{6\times 112}{10(10^2-1)}$

$= 1 - \frac{672}{10(100-1)} \quad = 1 - \frac{672}{10(99)} \quad = 1 - \frac{672}{990}$

$= 1 - 0.679 = 0.321$

∴ $r_{23} = +0.321 \Rightarrow$ Moderate degree $+ve$ correlation

Similarly, $r_{31} = 1 - \frac{6\sum {D_3}^2}{N(N^2-1)}$

$= 1 - \frac{6\times 182}{10(10^2-1)}$

$= 1 - \frac{1092}{10(100-1)} \quad = 1 - \frac{1092}{10(99)} \quad = 1 - \frac{1092}{990}$

$= 1 - 1.103 = -0.103$

∴ $r_{31} = -0.103 \Rightarrow$ Low degree $-ve$ correlation

⇒ Since r_{23} is highest, so $2nd$ and $3rd$ judges agree the most.

Also, r_{31} being lowest, $3rd$ and $1st$ judges disagree the most.

6.4.2 Spearman's Rank Correlation when ranks are not given.

1. Assign the ranks in descending order to series X by giving first rank to highest value and second rank to value lower than higher value and so on.
2. Similarly assign the ranks to series Y.
3. Calculate the difference between ranks of both the series denoted by ∑D.
4. Take square of deviations and calculate the value of D^2.
5. Calculate sum of square of deviations denoted by $\sum D^2$.
6. Apply following formula.

Example 11. ***Following are the marks obtained by* 8 *students in Maths and Statistics. Find the Rank Correlation Coefficient.***

Marks in Maths	**60**	**70**	**53**	**59**	**68**	**72**	**50**	**54**
Marks in Statistics	**44**	**74**	**54**	**64**	**84**	**79**	**53**	**66**

Solution:

X	Ranks R_1	Y	Ranks R_2	Difference of Ranks $D = R_1 - R_2$	D^2
60	4	44	8	-4	16
70	2	74	3	-1	1
53	7	54	6	$+1$	1
59	5	64	5	0	0
68	3	84	1	$+2$	4
72	1	79	2	-1	1
50	8	53	7	$+1$	1
54	6	66	4	$+2$	4
					$\sum D^2 = 28$

Here $N = 8$

⇒ Rank Coefficient of Correlation, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$$= 1 - \frac{6\times 28}{8(8^2-1)} \qquad = 1 - \frac{168}{8(64-1)}$$

$$= 1 - \frac{168}{8(63)} \qquad = 1 - \frac{168}{504}$$

$$= 1 - 0.33 = 0.67$$

⇒ Rank Correlation Coefficient $= 0.67$

6.5.3 Spearman’s Rank Correlation when there is repetition in ranks.

1. Assign the ranks in descending order to series X by giving first rank to highest value and second rank to value lower than higher value and so on. If two items have same value, assign the average rank to both the item. For example, two equal values have ranked at 5th place than rank to be given is 5.5 to both i.e. mean of 5th and 6th rank. **($\frac{5+6}{2}$).**
2. Similarly assign the ranks to series Y.
3. Calculate the difference between ranks of both the series denoted by ∑D.
4. Take square of deviations and calculate the value of D^2.
5. Calculate sum of square of deviations denoted by $\sum D^2$.
6. Apply following formula.

$$\boldsymbol{r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}(m_1{}^3 - m_1) + \frac{1}{12}(m_2{}^3 - m_2)\right\}}{N(N^2 - 1)}}$$

Where m = no. of times a particular item is repeated.

Example 12. ***Find the Spearsman’s Correlation Coefficient for the data given below***

$\boldsymbol{X}$	**110**	**104**	**107**	**82**	**93**	**93**	**115**	**95**	**93**	**113**
$\boldsymbol{Y}$	**80**	**78**	**90**	**75**	**81**	**70**	**87**	**78**	**73**	**85**

Solution: Here, in X series the value 93 occurs thrice $(m_1 = 3)$, $i.e.$ at 7^{th}, 8^{th} and 9^{th} rank. So all the three values are given the same average rank, $i.e. \frac{7+8+9}{3} = 8^{th}$ rank.

Similarly, in Y series the value 78 occurs twice $(m_2 = 2)$, $i.e.$ at 6^{th} and 7^{th} rank. So both the values are given the same average rank, $i.e. \frac{6+7}{2} = 6.5^{th}$ rank.

X	Ranking of X R_1	Y	Ranking of Y R_2	Difference of Ranks $D = R_1 - R_2$	D^2
110	3	80	5	−2	4
104	5	78	6.5	−1.5	2.25
107	4	90	1	+3	9
82	10	75	8	+2	4
93	8	81	4	+4	16
93	8	70	10	−2	4
115	1	87	2	−2	1
95	6	78	6.5	−0.5	0.25
93	8	73	9	−1	1
113	2	85	3	−1	1
					$\sum D^2 = 42.5$

Here $N = 10$

Spearman’s Rank Correlation Coefficient, $r = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}({m_1}^3 - m_1) + \frac{1}{12}({m_2}^3 - m_2)\right\}}{N(N^2-1)}$

$i.e.$ $$r = 1 - \frac{6\left\{42.50 + \frac{1}{12}(3^3-3) + \frac{1}{12}(2^3-2)\right\}}{10(10^2-1)}$$

$$= 1 - \frac{6\left\{42.50 + \frac{24}{12} + \frac{6}{12}\right\}}{10(100-1)}$$

$$= 1 - \frac{6\left\{42.50 + 2 + \frac{1}{2}\right\}}{10 \times 99} \qquad = 1 - \frac{6\{42.5+2.5\}}{990} \qquad = 1 - \frac{6 \times 45}{990}$$

$$= 1 - 0.2727 = 0.7273$$

$\Rightarrow$ Rank Correlation Coefficient $= 0.7273$

Example 13. ***The rank correlation coefficient between the marks obtained by ten students in Mathematics and Statistics was found to be $\mathbf{0.5}$. But later on, it was found that the difference in ranks in the two subjects obtained by one students was wrongly taken as $\mathbf{6}$ instead of $\mathbf{9}$. Find the correct rank correlation.***

Solution : Given $N = 10$, Incorrect $r = 0.5$

We know that

Rank Correlation Coefficient, $r = 1 - \frac{6\sum D^2}{N(N^2-1)}$

$$\Rightarrow \qquad 0.5 = 1 - \frac{6\sum D^2}{10(10^2-1)} = 1 - \frac{6\sum D^2}{10 \times 99}$$

$\Rightarrow$ Incorrect $\sum D^2 = \frac{990}{6} \times 0.5 = 82.5$

$\therefore$ The corrected value of $\sum D^2 = 82.5 - 6^2 + 9^2$

$$= 82.5 - 36 + 81 = 127.5$$

$\therefore$ Correct Rank Correlation Coefficient, $r = 1 - \frac{6 \times 127.5}{10(10^2-1)}$

$$= 1 - \frac{765}{10(100-99)}$$

$$= 1 - \frac{765}{10 \times 99} \qquad = 1 - \frac{765}{990}$$

$$= 1 - 0.7727$$

$$= 0.2273$$

TEST YOUR UNDERSTANDING (A)

1. Find Rank correlation on base of following data.

X	78	36	98	25	75	82	90	62	65	39
Y	84	51	91	60	68	62	86	58	53	47

In Dance competition following ranks were given by 3 judges to participants. Determine which two judges have same preference for music:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

3. Find Rank correlation on base of following data.

X	25	30	38	22	50	70	30	90
Y	50	40	60	40	30	20	40	70

4. Find Rank correlation on base of following data.

X	63	67	64	68	62	66	68	67	69	71
Y	66	68	65	69	66	65	68	69	71	70

Answers

1. .82	1. 00
2. I and II -.2121, II and III -.297, I and III .6364, so judge I and III	2. 0.81

6.6 MEANING OF REGRESSION ANALYSIS

According to Sir Francis Galton, the term regression analysis is defined as "the law of regression that tells heavily against the full hereditary transmission of any gift, the more bountifully the parent is gifted by nature, the rarer will be his good fortune if he begets a son who is richly endowed as himself, and still more so if he has a son who is endowed yet more largely."

In the words of Ya Lun Chou, "Regression analysis attempts to establish the nature of the relationship between variables that is to study the functional relationship between the variables and thereby provide mechanism for prediction or forecasting".

6.7 DIFFERENT TYPES OF REGRESSION ANALYSIS

1. **Simple Regression**: When there are only two variables under study it is known as a simple regression. For example, we are studying the relation between Sales and Advertising expenditure. If we consider sales as Variable X and advertising as variable Y, then the X = a+b Y is known as the regression equation of X on Y where X is the dependent variables and Y and the independent variable. In other words, we can find the value of variable X (Sales) if the value of Variable Y (Advertising) is given.
2. **Multiple Regression:** The study of more than two variables at a time is known as multiple regression. Under this, only one variable is taken as a dependent variable and all the other

variables are taken as independent variables. For example, If we consider sales as Variable X, advertising as variable Y and Income as Variable Z, then using the functional relation X = *f* (Y, Z), we can find the value of variable X (Sales) if the value of Variable Y (Advertising) and the value of variable Z (Income) is given.

3. **Total Regression:** Total regression analysis is one in which we study the effect of all the variables simultaneously. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we can study impact of advertising and income simultaneously on sales. This is a case of total regression analysis. In such cases, the regression equation is represented as follows:

   X = *f* (Y, Z),

4. **Partial Regression:** In the case of Partial Regression one or two variables are taken into consideration and the others are excluded. For example, when we want to study the effect of advertising expenditure of business represented by variable Y, income of the consumer represented by variable Z, on the amount of sales represented by variable X, we will not study impact of both income and advertising simultaneously, rather we will first study effect of income on sales keeping advertising constant and then effect of advertising on sales keeping income constant. Partial regression can be written as

   X=f (Y not Z).

5. **Linear Regression:** When the functional relationship between X and Y is expressed as the first-degree equations, it is known as linear regression. In other words, when the points plotted on a scatter diagram concentrate around a straight line it is the case of linear regression.

6. **Non-linear Regression**: On the other hand, if the line of regression (in scatter diagram) is not a straight line, the regression is termed as curved or non-linear regression. The regression equations of non-linear regression are represented by equations of higher degree. The following diagrams show the linear and non-linear regressions:

6.8 PROPERTIES OF REGRESSION COEFFICIENTS

The regression coefficients discussed above have a number of properties which are given as under:

1. The geometric Mean of the two regression coefficients gives the coefficients of correlation i.e. r = $\sqrt{bxy * byx}$
2. Both the regression coefficients must have the same sign i.e., in other words either both coefficients will have + signs or both coefficients will have - signs. This is due to the fact that in first property we have studied that geometric means of both coefficients will give us value of correlation. If one sign will be positive and other will be negative, the product of both signs will be negative. And it is not possible to find out correlation of negative value.
3. The signs of regression coefficients will give us signs of coefficient of correlation. This means if the regression coefficients are positive the correlation coefficient will be positive, and if the regression coefficients are negative then the correlation coefficient will be negative.
4. If one of the regression coefficients is greater than unity or 1, the other must be less than unity. This is because the value of coefficient of correlation must be in between ± 1. If both the regression coefficients are more than 1, then their geometric mean will be more than 1 but the value of correlation cannot exceed 1.
5. The arithmetic mean of the regression coefficients is either equal to or more than the correlation coefficient $\frac{\text{bxy+byx}}{2} \geq \sqrt{bxy * byx}$
6. If the regression coefficients are given, we can calculate the value of standard deviation by using the following formula.

   a. bxy = $r\,\frac{\sigma x}{\sigma y}$ or byx = $r\,\frac{\sigma y}{\sigma x}$
7. Regression coefficients are independent of change of origin but not of scale. This means that if the original values of the two variables are added or subtracted by some constant, the values of the regression coefficients will remain the same. But if the original values of the two variables are multiplied, or divided by some constant (common factors) the values of the regression equation will not remain the same.

6.9 RELATIONSHIP BETWEEN CORRELATION AND REGRESSION

1. Correlation is a quantitative tool that measure of the degree of relationship that is present between two variables. It shows the degree and direction of the relation between tow variables. Regression helps us to find the value of a dependent variable when the value of independent variable is given.

2. Correlation between two variables is the same. For example, we calculate the correlation between sales and advertising or advertising and sales, the value of correlation will remain same. But this is not true for Regression. Regression equation of Advertising on sales will be different from regression equation of Sales on advertising.
3. If there is positive correlation, the distance between the two lines will be less. That means the two regression lines will be closer to each other- Similarly, if there is low correlation, the lines will be farther to each other. A positive correlation implies that the lines will be upward sloping whereas a negative correlation implies that the regression lines will be downward sloping.
4. Correlation between two variables can be calculated by taking the square root of the product of the two regression coefficients.

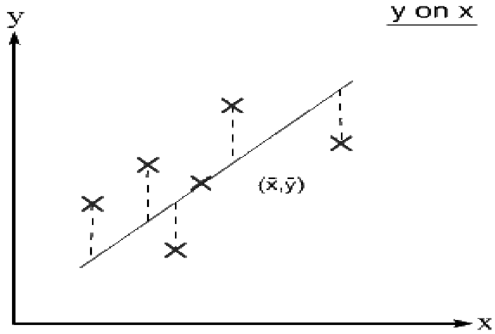
Following are some of the differences between Correlation and Regression:

1. Correlation measures the degree and direction of relationship between two variables. Regression measures the change in value of a dependent variable given the change in value of an independent variable.
2. Correlation does not depict a cause and effect' relationship. Regression depicts the causal relationship between two variables.
3. Correlation is a relative measure of linear relationship that exists between two variables. Regression is an absolute measure which measures the change in value of a variable.
4. Correlation between two variables is the same. In other words, Correlation between two variables is the same. rxy =ryx. Regression is not symmetrical in formation. So, the regression coefficients of X on Y and of Y on X are different.
5. Correlation is independent of Change in origin or scale. Regression is independent of Change in origin but not of scale.
6. Correlation is not capable of any further mathematical treatment. Regression can be further treated mathematically.
7. Coefficient of correlation always lies between -1 and +1. Regression coefficient can have any value.

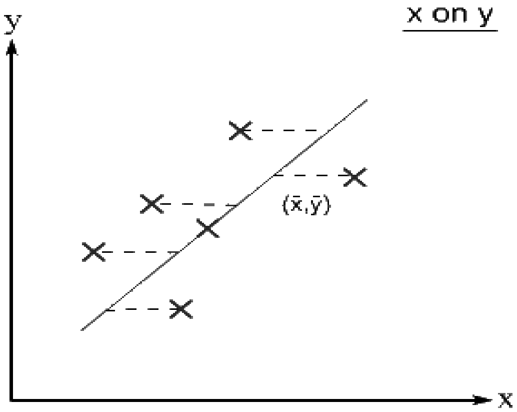
6.10 MEANING OF REGRESSION LINES

The lines that are used in Regression for the purpose of estimation are called as regression line. In other words, the lines that are used to study the dependence of one variable on the other variable are called as regression line. If we have two variables X and Y then there.

a. Regression Line of Y on X: Regression Line Y on X measures the dependence of Y on X and we can estimate the value of Y for the given values of X. In this line Y is dependent variable and X is independent variable.

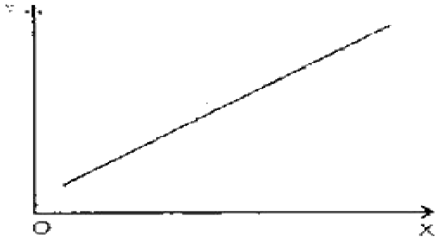


b. Regression Line of X on Y: Regression Line X on Y measures the dependence of X on Y and we can estimate the value of X for the given values of Y. In this line X is dependent variable and Y is independent variable.



The direction of two regression equation depends upon the degree of correlation between two variables. Following can be the cases of correlation between two variables:

1. Perfect positive correlation: If there is a perfect positive correlation between two variables (i.e., r = +1), both the lines will coincide with each other and will be having



Both the line will coincide with positive slope positive slope. Both the lines X on Y and Yon X will be same in this case. In other words, in that case only one regression line can be drawn as shown in the diagram. The slope of the line will be upward.

2. Perfect negative correlation: If there is a perfect negative correlation between two variable (i.e. r = -1), both the lines will coincide with each other and will in such case these lines will be having negative slope. Both the lines X on Y and Yon X will be same but downward sloping. In other words, in that case only one regression line can be drawn as shown in the diagram. The slope of the line will be upward.

3. High degree of correlation: If there is a high degree of correlation between two variables, both the lines will be near to each other. In other words, these lines will be closer to each other but the lines will not coincide with each other. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.

4. Low degree of correlation: If there is a low degree of correlation between two variables, both the lines will be having more distance from each other. In other words, these lines will be farther to each other, that is the gap between the two lines will be more. Both the lines will be separate. Further the direction of lines depends upon the positive or negative correlation.

5. No correlation: If there is a no correlation between two variables (i.e., r = 0), both the lines will be perpendicular to each other. In other words, these lines will cut each other at $90^{0.}$ This diagram depicts the perpendicular relation between the two regression lines when there is absolutely zero correlation between the two variables under the study.

6.11 LEAST SQUARE METHOD OF FITTING REGRESSION LINES

Under this method the lines of best fit is drawn as the lines of regression. These lines of regression are known as the lines of the best fit because, with help of these lines we can make the estimate of the values of one variable depending on the value of another variable. According to the Least Square method, regression line should be plotted in such a way that sum of square of the difference between actual value and estimated value of the dependent variable should be least or minimum possible. Under this method we draw two regression lines that are

a. **Regression line Y on X** – it measures the value of Y when value of X is given. In other words, it assumes that X is an independent variable whereas the other variable Y is dependent variable. Mathematically this line is represented by

$$Y = a + bX$$

Where Y – Dependent Variable

X – Independent Variable

a & b – Constants

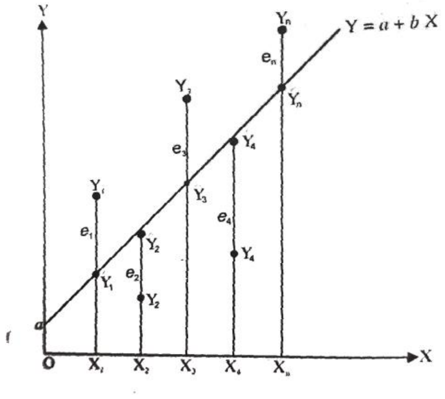
b. **Regression line X on Y** – it measures the value of X when value of Y is given. In other words it assumes that Y is an independent variable whereas the other variable X is dependent variable. Mathematically this line is represented by

$$X = a + bY$$

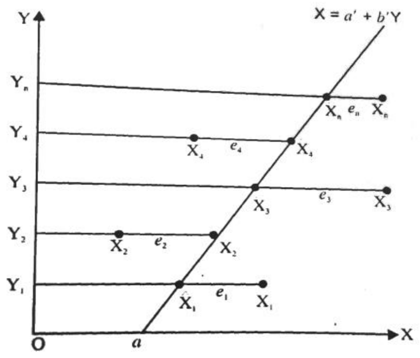
Where X – Dependent Variable

Y – Independent Variable

a & b – Constants



Equation Y on X



Equation X on Y

In the above two regression lines, there are two constants represented by "a" and "b". The constant "b" is also known as regression coefficient, which are denoted as "byx" and "bxy", Where "byx" represent regression coefficient of equation Y on X and "bxy" represent regression coefficient of equation X on Y. When the value of these two variables "a" and "b" is determined we can find out the regression line.

6.11.1 DIRECT METHODS TO ESTIMATE REGRESSION EQUATION

The regression equations can be obtained by 'Normal Equation Method" as follows:

1. **Regression Equation of Y on X:** The regression equation Y on X is in the format of Y = a + bX, where Y is a Dependent Variable and X is an Independent Variable. To estimate this regression equation, following normal equations are used:

   $\Sigma Y = na + b_{yx} \Sigma X$

   $\Sigma XY = a \Sigma X + b_{yx} \Sigma X^2$

   With the help of these two equations the values of ‘a' and 'b' are obtained and by putting the values of ‘a' and 'b' in the equation Y = a + b X we can predict or estimate value of Y for any value of X.

2. **Regression Equation of X on Y:** The regression equation X on Y is in the format of X = a + bY, where X is a Dependent Variable and Y is an Independent Variable. To estimate this regression equation, following normal equations are used:

   $\Sigma X = na + b_{xy} \Sigma Y$

   $\Sigma XY = a \Sigma Y + b_{xy} \Sigma Y^2$

   With the help of these two equations the values of ‘a' and 'b' are obtained and by putting the values of ‘a' and 'b' in the equation X = a + bY we can predict or estimate value of Y for any value of X.

Example 1. Find out the two regression lines for the data given below using the method of least square.

Variable X:	5	10	15	20	25
Variable Y:	20	40	30	60	50

Determination of the regression lines by the method of least square. Also find out

a. Value of Y when value of X is 40
b. Value of X when value of Y is 80.

Solution:

X	**Y**	**X2**	**Y2**	**XY**
5	20	25	400	100
10	40	100	1600	400
15	30	225	900	450
20	60	400	3600	1200

25	50	625	2500	1250
XX = 75	XY = 200	XX2=1375	XY2=9,000	XXY =3400

(i) Regression line of Y on X

This is given by Y = *a + bX*

where *a* and *b* are the two constants which are found by solving simultaneously the two normal equations as follows:

$\Sigma Y = na + b_{yx}\Sigma X$

$\Sigma XY = a\,\Sigma X + b_{yx}\Sigma X^2$

Substituting the given values in the above equations we get,

$200 = 5a + 75b$ …………………………………….. (i)

$3400 = 75a + 1375b$ …………………………………….. (ii)

Multiplying the eqn. (i) by 15 we get

$3000 = 75a + 1125b$…………………………………….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

$3400 = 75a + 1375b$

$\underline{-3000 = -75a - 1125b}$

$400 = 250b$

or $b = 1.6$

Putting the above value of b in the eqn. (i) we get,

$200 = 5a + 75(1.6)$ or

$5a = 200 - 120$ or

$a = 16$

Thus, a = 16, and b = 1.6

Putting these values in the equation Y = *a* + bX we get

Y = 16+ 1.6X

So, when X is 40, the value of Y will be

$Y = 16 + 1.6(40) = 80$

(ii) Regression line of X on Y

This is given by X = *a + bY*

where *a* and *b* are the two constants which are found by solving simultaneously the two normal equations as follows:

$\Sigma X = na + b_{xy}\Sigma\ Y$

$\Sigma XY = a\ \Sigma\ Y + b_{xy}\Sigma\ Y^2$

Substituting the given values in the above equations we get,

75 = 5a+200b ……….………………………………….. (i)

3400 = 200a + 9000b …………………………………….. (ii)

Multiplying the eqn. (i) by 40 we get

3000 = 200a + 8000b…………………………………….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

3400 = 200a + 9000b

<u>-3000 = -200a + -8000b</u>

400 = 1000b

or b = .4

Putting the above value of b in the eqn. (i) we get,

75 = 5a + 200(.4) or

5a = -5 or

a = -1

Thus, a = -1, and b = .4

Putting these values in the equation X = *a + bY* we get

X = -1+ .4Y

So when Y is 80, the value of X will be

X = -1+ .4(80) = 31

6.11.2 OTHER METHODS OF ESTIMATING REGRESSION EQUATION

This method discussed above is known as direct method. This is one of the popular methods of finding the regression equation. But sometime this method of finding regression equation becomes cumbersome and lengthy specially when the values of X and Y are very large. In this case we can simplify the calculation by take the deviations of X and Y than dealing with actual values of X and Y. In such case

Regression equation Y on X

Y = a + bX

will be converted to $(Y - \overline{Y}) = byx\ (X - \overline{X})$

Similarly, Regression equation X on Y:

X = a + bY

will be converted into $(X - \overline{X}) = bxy\ (Y - \overline{Y})$

Now when we are using these regression equations, the calculations will become very simple as now we have to calculate value of only one constant that is value of "b" which is our regression coefficient. As there are two regression equations, so we need to calculate two regression coefficients that is Regression Coefficient X on Y, which is symbolically denoted as "bxy" and similarly Regression Coefficient Y on X, which is denoted as "byx". However, these coefficients can also be calculated using different methods. As we take deviations under this method, we can take deviations using actual mean, assumed mean or we can calculate it by not taking the deviations. Following formulas are used in such cases:

Method	**Regression Coefficient X on Y**	**Regression Coefficient Y on X**
When deviations are taken from actual mean	$bxy = \frac{\sum xy}{\sum y^2}$	$byx = \frac{\sum xy}{\sum x^2}$
When deviations are taken from assumed mean	$bxy = \frac{N\sum dxdy - \sum dx\sum dy}{N\sum dy^2 - (\sum dy)^2}$	$byx = \frac{N\sum dxdy - \sum dx\sum dy}{N\sum dx^2 - (\sum dx)^2}$
Direct Method: Using sum of X and Y	$bxy = \frac{N\sum XY - \sum X\sum Y}{N\sum Y^2 - (\sum Y)^2}$	$byx = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - (\sum X)^2}$
Using the correlation coefficient (r) and standard deviation (σ)	$bxy = r \cdot \frac{\sigma x}{\sigma y}$	$byx = r \cdot \frac{\sigma y}{\sigma x}$

Example 2. From the information give below obtain two regression lines X on Y and Yon X using

1. Actual Mean Method.
2. Assumed Mean Method
3. Direct Method (Without taking Mean)

Number of Hrs Machine Operated	7	8	6	9	11	9	10	12
Production (Units in 000):	4	5	2	6	9	5	7	10

Solution:

1. **Actual Mean Method**

Calculation of Regression Equation

X	Y	x =X - $\bar{X}$	x^2	y =Y - $\bar{Y}$	y^2	xy
7	4	-2	4	-2	4	4
8	5	-1	1	-1	1	1
6	2	-3	9	-4	16	12
9	6	0	0	0	0	0
11	9	2	4	3	9	6
9	5	0	0	-1	1	0
10	7	1	1	1	1	1
12	10	3	9	4	16	12
$\sum X = 72$	$\sum Y = 48$		$\sum x^2 = 28$		$\sum y^2 = 48$	$\sum xy = 36$

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where bxy = $\frac{\sum xy}{y^2}$

$= \frac{36}{48}$

$= .75$

So $(X - 9) = .75 (Y - 6)$

$X - 9 = .75Y - 4.5$

X = 4.5 + .75Y

Regression equation of Y on X:

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

Where byx = $\frac{\sum xy}{x^2}$

$= \frac{36}{28}$

$= 1.286$

So $(Y - 6) = 1.286 (X - 9)$

$Y - \ 6 = 1.286X - 11.57$

$Y = -5.57 + 1.286X$

2. <u>Assumed Mean Method</u>

Calculation of Regression Equation

X	Y	dx =X – A (A = 8)	dx^2	dy =Y – A (A = 5)	dy^2	dxdy
7	4	-1	1	-1	1	1
8	5	0	0	0	0	0
6	2	-2	4	-3	9	6
9	6	1	1	1	1	1
11	9	3	9	4	16	12
9	5	1	1	0	0	0
10	7	2	4	2	4	4
12	10	4	16	5	25	20
$\sum X = 72$	**$\sum Y = 48$**	**$\sum dx = 8$**	**$\sum dx^2 = 36$**	**$\sum dy = 8$**	**$\sum dy^2 = 56$**	**$\sum xy = 44$**

$\overline{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\overline{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

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

$= \frac{8\,(44) - (8)\,(8)}{8\,(56) - (8)^2}$

$= \frac{352 - 64}{448 - 64}$

$= \frac{288}{384}$

$= .75$

So $(X - \ 9) = .75\,(Y - \ 6)$

$X - \ 9 = .75Y - 4.5$

$X = 4.5 + .75Y$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{8\ (44) - (8)\ (8)}{8\ (36) - (8)^2} \quad = \quad \frac{288}{224} \quad = 1.286$

So (Y – 6) = 1.286 (X – 9)

Y – 6 = 1.286X – 11.57

Y = – 5.57+ 1.286X

3. <u>Direct Method (Without taking Mean)</u>

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
7	4	49	16	28
8	5	64	25	40
6	2	36	4	12
9	6	81	36	54
11	9	121	81	99
9	5	81	25	45
10	7	100	49	70
12	10	144	100	120
$\sum$X =72	**$\sum$Y =48**	**$\sum X^2$ = 676**	**$\sum Y^2$ = 336**	**$\sum$XY = 468**

$\bar{X} = \frac{\sum X}{N} = \frac{72}{8} = 9$

$\bar{Y} = \frac{\sum Y}{N} = \frac{48}{8} = 6$

Regression equation of X on Y:

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

Where bxy = $\frac{N\sum XY - \sum X \sum Y}{N\sum Y^2 - (\sum Y)^2}$

$= \frac{8\ (468) - (72)\ (48)}{8\ (336) - (48)^2} \quad = \frac{3744 - 3456}{2688 - 2304}$

$= \frac{288}{384}$

= .75

So (X – 9) = .75 (Y – 6)

X – 9 = .75Y – 4.5

X = 4.5 + .75Y

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$

$= \frac{8\,(468) - (72)\,(48)}{8\,(676) - (72)^2}$

$= \frac{3744 - 3456}{5408 - 5184} \quad = \frac{288}{224}$

= 1.286

So (Y – 6) = 1.286 (X – 9)

Y – 6 = 1.286X – 11.57

Y = – 5.57+ 1.286X

Example 3. Find out two Regression equations on basis of the data given below:

	X	Y
Mean	60	80
Standard Deviation (S.D.)	16	20
Coefficient of Correlation	.9	

Also find value of X when Y = 150 and value of Y when X = 100.

Solution: Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

Where $bxy = r\,\frac{\sigma x}{\sigma y}$

$= .9\ \frac{16}{20} \quad = .72$

So (X – 60) = .72 (Y – 80)

X – 60 = .72Y – 57.6

X = 2.4 + .72Y

When Y = 150 than X = 2.4 + .72(150) = 110.4

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = r\,\frac{\sigma y}{\sigma x}$

$= .9\ \frac{20}{16} \quad = 1.125$

So $(Y - 80) = 1.125 (X - 60)$

$Y - 80 = 1.125X - 67.5$

$Y = 12.5 + 1.125 X$

When X = 100 than $Y = 12.5 + 1.125 (100) = 125$

Example 4. From the following data find out two lines of regression land also find out value of correlation.

$\sum X = 250$; $\sum Y = 300$; $\sum XY = 7900$;

$\sum X^2 = 6500$; $\sum Y^2 = 10000$; $n = 10$

Solution:

$$\overline{X} = \frac{\sum X}{N} = \frac{250}{10} = 25$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{300}{10} = 30$$

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

Where $byx = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - (\sum X)^2}$

$$= \frac{10 (7900) - (250) (300)}{10 (6500) - (250)^2}$$

$$= \frac{79000 - 75000}{65000 - 62500}$$

$$= \frac{4000}{2500}$$

$= 1.6$

So $(Y - 30) = 1.6 (X - 25)$

$Y - 30 = 1.6X - 40$

$Y = -10 + 1.6 X$

Regression equation of X on Y:

$(X - \overline{X}) = b_{xy} (Y - \overline{Y})$

Where $bxy = \frac{N\sum XY - \sum X\sum Y}{N\sum Y^2 - (\sum Y)^2}$

$$= \frac{10 (7900) - (250) (300)}{10 (10000) - (300)^2} = \frac{79000 - 75000}{100000 - 90000} = \frac{4000}{10000} = 0.4$$

So $(X - 25) = .4 (Y - 30)$

$X - \ 25 = .4Y - 12$

X = 13 + .4Y

Coefficients of Correlation

$r = \sqrt{bxy\ *\ byx}$

$r = \sqrt{1.6\ *\ 0.4}$

$r = \sqrt{.64}$

r = .8

Example 5. From the following data find out two lines of regression land also find out value of correlation. Also find value of Y when X = 30

$\sum X = 140$;	$\sum Y = 150$;	$\sum (X - 10)(Y - 15) = 6$;
$\sum (X - 10)^2 = 180$;	$\sum (Y - 15)^2 = 215$;	n = 10

Solution:

Let's take assumed mean of Series X = 10 and Series Y = 15.

$\sum dx = \sum (X - 10) = \sum X - 10n = 140 - 100 = 40$

$\sum dy = \sum (Y - 15) = \sum Y - 15n = 150 - 150 = 0$

$\sum dx^2 = \sum (X - 10)^2 = 180$

$\sum dy^2 = \sum (Y - 15)^2 = 215$

$\sum dxdy = \sum (X - 10)(Y - 15) = 6$

So

$\overline{X} = A + \frac{\sum X}{N} = 10 + \frac{40}{10} = 14$

$\overline{Y} = A + \frac{\sum Y}{N} = 15 + \frac{0}{10} = 15$

Regression equation of Y on X:

$(Y - \ \overline{Y}) = b_{xy} (X - \ \overline{X})$

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

$= \frac{10(6) - (40)\ (0)}{10\ (180) - (40)^2} \qquad = \frac{60}{200} \qquad = .3$

So $(Y - \ 15) = .3\ (X - \ 14)$

$Y - \ 15 = .3X - 4.2$

Y = 10.8+ .3X

When X = 30 than Y = 10.8 + .3(30) = 19.8

Regression equation of Y on X:

$(Y - \overline{Y}) = b_{xy} (X - \overline{X})$

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

$= \frac{10\ (6) - (40)\ (0)}{10\ (25) - (0)^2} \quad = \quad \frac{60}{250} \quad = .24$

So $(Y - 15) = .24 (X - 14)$

$Y - 15 = .24X - 3.36$

Y = 11.64+ .24X

Coefficients of Correlation

$r = \sqrt{bxy * byx} = \sqrt{.3 * .24}$

$r = \sqrt{.072}$

r = .268

Example 5. From the following data find out which equation is equation X on Y and which equation is equation Y on X. Also find $\overline{X}$, $\overline{Y}$ and r.

$3X + 2Y - 26 = 0$

$6X + Y - 31 = 0$

Solution:

To find $\overline{X}$ and $\overline{Y}$, we will solve following simultaneous equations

3X + 2Y = 26 …………………………….. (i)

6X + Y = 31 ……………………………. (ii)

Multiply equation (i) with 2, we get

6X + 4Y = 52 …………………………….. (iii)

Deduct equation (ii) from equation (iii)

$$\begin{array}{r} 6X + 4Y = 52 \\ -\underline{6X - Y = -31} \\ 3Y = 21 \end{array}$$

Y = 7

Or $\overline{Y}$ = 7.

Put the value of Y in Equation (i), we get

3X + 2(7) = 26

3X + 14 = 26

3X = 12

X = 4

or $\overline{X}$ = 4

Let 3X + 2Y = 26 be regression equation X on Y

3X = 26 – 2Y

X = $\frac{26}{3} - \frac{2}{3}$ Y

So $b_{xy} = -\frac{2}{3}$

Let 6X + Y = 31 be regression equation Y on X

Y = 31 – 6X

So b_{yx} = – 6

As r = $\sqrt{bxy \ * \ byx}$

r = $-\sqrt{-(\frac{2}{3}) \times (-.6 \ }\ )$

r = -2, but this is not possible as value of r always lies between $-1\, and + 1$. So, our assumption is wrong and equation are reverse.

Let 6 X + Y = 31 be regression equation X on Y

6X = 31 –Y

X = $\frac{31}{6} - \frac{1}{6}$ Y

So $b_{xy} = -\frac{1}{6}$

Let 3X + 2Y = 26 be regression equation Y on X

2Y = 26 – 3X

Y = $\frac{26}{2} - \frac{3}{2}$ X

So $b_{yx} = -\frac{3}{2}$

As r = $\sqrt{bxy \ * \ byx}$

r = $-\sqrt{-(\frac{1}{6}) \times -(\frac{3}{2}) \ }\ )$

r = $-.5$, which is possible. So, our assumption is right.

So,

$\overline{Y} = 7$; $\overline{X} = 4$;

X on Y is X = $\frac{31}{6} - \frac{1}{6}$ Y

Y on X is Y = $\frac{26}{2} - \frac{3}{2}$ X

r = $-.5$

TEST YOUR UNDERSTANDING

1. Find both regression equations:

X	6	2	10	4	8
Y	9	11	5	8	7

2. From following estimate the value of Y when X = 30 using regression equation.

X	25	22	28	26	35	20	22	40	20	18	19	25
Y	18	15	20	17	22	14	15	21	15	14	16	17

3. Fit two regression lines:

X	30	32	38	35	40
Y	10	14	16	20	15

Find X when Y = 25 and find Y when X = 36.

4. Find out two Regression equations on basis of the data given below:

	X	Y
Mean	65	67
Standard Deviation (S.D.)	2.5	3.5
Coefficient of Correlation	.8	

5. In a data the Mean values of X and Y are 20 and 45 respectively. Regression coefficient b_{yx} = 4 and b_{xy} 1/9. Find

a. coefficient of correlation
b. Standard Deviation of X, if S.D. of Y = 12
c. Find two regression lines

6. You are supplied with the following information. Variance of X = 36

12X – 51Y + 99 = 0

60X – 27Y = 321.

Calculate

(a) The average values of X and Y

(b) The standard deviation of Y and

7. The lines of regression of Y on X and X on Y are Y = X + 5 and 16X = 9Y + 4 respectively

Also σy=4 Find $\bar{X}$, $\bar{Y}$, σx and r.

8. Given:

$$\sum X = 56 , \sum Y = 40 , \sum X^2 = 524$$
$$\sum Y^2 = 256 , \sum XY = 364 , N = 8$$

(a) $find\ the\ regression\ equatiopn\ of\ X\ on\ Y$

Answers

1. X = 16.4 – 1.3Y, Y = 11.9 - .65	4. Y = 1.12X – 5.8, X = .57Y +26,81	7. Mean of X = 7, Mean of Y= 12, S.D of X = 3, r= .75,
2. 18.875	5. .67, 2, Y= 4X – 35 and X = 1/9 Y +15	8. X = 1.5Y - 0.5, r = .977
3. Y = .46X – 1.1, X = .6Y + 26, Value of Y = 15.46, Value of X = 40.25	6. Mean of X = 13, Mean of Y= 17, S.D of Y = 8	

6.12 LET US SUM UP

- Correlation shows the relation between two or more variables.
- Value of the coefficient of correlation always lies between -1 and +1.
- Correlation may be positive or negative.
- Correlation may be linear or nonlinear.
- Karl Person’s coefficient of correlation is the most popular method of correlation.
- Spearman’s Rank correlation calculated correlation on the basis of ranks given to data.
- It can deal with qualitative data also.
- Regression is a useful tool of forecasting.

- With help of regression, we can predict the value of can find the value of X if value of Y is given or value of Y if value of X is given.
- It creates the mathematical linear relation between two variables X and Y, out of which one variable is dependent and other is independent.
- In this we find out two regression equations.
- Regression can be linear or nonlinear.
- It can be simple or multiple.
- Regression is based on the principle of Least Squares.
- Correlation coefficient can find out with help of regression coefficients.

6.13 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Define the following

Q1. Correlation

Q2. List the types of correlation

Q3. Formula of Karl Pearson correlation

Q4. Formula of repeated rank

Q5. Regression

Q6. Types of regression

B. Long Answer Type Questions

Q1. What is Correlation? What are uses of measuring correlation.

Q2. Give different types of correlation.

Q3. Give Karl Persons method of calculating correlation.

Q4. Give Karl Pearson’s coefficient of correlation in case of actual and assumed mean.

Q5. What are merits and limitations of Karl Pearson’s method?

Q6. What is Spearman’s Rank correlation? How it is determined.

Q7. What is Regression.? Explain its uses.

Q8. What is relation between Regression and correlation?

Q9. Explain different types of regressions.

Q10. How two regression lines are determined under direct method?

Q11. Explain various methods of finding regression equations.

Q12. What are properties of regression coefficients?

6.14 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics*, Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics*, Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi.
- M.R. Spiegel, *Theory and Problems of Statistics,* Schaum's Outlines Series, McGraw Hill Publishing Co.
- Richard Levin and David S. Rubin, *Statistics for Management*, Prentice Hall of India, New Delhi. Hill Publishing Co.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 7: TIME SERIES ANALYSIS AND INDEX NUMBER

STRUCTURE

7.0 LEARNING OBJECTIVES

After studying the Unit, students will be able to:

- Define the meaning of time series Analysis and index numbers
- Distinguish different types of fluctuations in the time series analysis
- Understand how time series analysis is useful for forecasting
- Apply various methods of time series in the prediction of trends
- Describe the uses of index numbers
- Understand how index numbers are prepared
- Understand uses of index numbers
- Construction of simple index numbers

- Test of consistency of index number

7.1 INTRODUCTION OF TIME SERIES

One of the important functions of business managers is to make forecasts about the future. This forecasting helps them in making the business decisions. There are many Statistical Techniques that help a business manager in forecasting the future. Time series analysis is one such technique. This technique is not only used by Business managers; other persons interested in forecasting also use this technique like economists, etc. Time series analysis is a tool with help of which we try to predict future values based on data available to us. For example, if we have data on sales of a company for last 10-12 years and we want to predict the likely sales of the company for the next year, we can do so using time series analysis. Following are few examples of time series analysis:

- A series of data related to production of goods, prices of goods or consumption level of goods.
- Data related to the rainfall or temperature of a region.
- The data related to sales profit etc. of any business firm.

In time series we collect the data related to statistical observations and place such data in chronological order, that means in the order of occurrence of these observations. On the basis of these observations, we can try to predict the future values of the observation. Following is the definition of time series analysis:

7.2 ESSENTIAL CONDITIONS OF TIME SERIES ANALYSIS

1. Time series analysis must consist of those values that are homogenous for example the sales data of every year must be in same quantities as in kilograms. If sales of some years are given in quantity and others are given in value, then we cannot apply time series analysis.
2. The data present must be about time only. So, out of two variables given, one variable should be time. For example, if a relation between Price and Demand is given it is not time series.
3. The data must be arranged in chronological order.
4. The data must be available for long period of time at least 10 to 12 years.
5. We must try to keep an equal gap between two periods.
6. If the gap between the periods is not equal and some values are missing, we should try to find out those values using the interpolation.
7. The data must have some relation with the time. For example, if we are measuring average marks of the students nn a class, it is not related to time.

7.3 ADVANTAGES OF TIME SERIES ANALYSIS

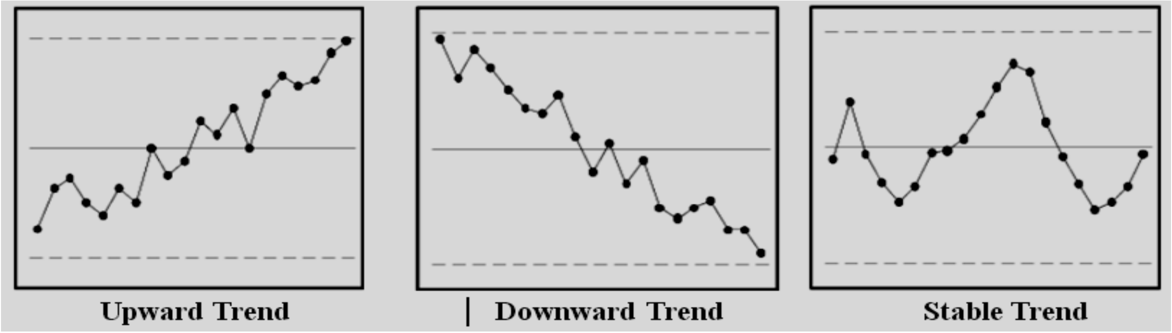
1. Time series analysis helps us in understanding the past behaviour of the phenomenon.
2. It helps us in predicting the future course of action.
3. It helps us in understanding how values change with the passage of time.
4. With help of time series, we can isolate impact of various factors like seasonal factors, cyclical factors or other irregular changes in the data.
5. With help of time series, we can find deviations between the actual achievements and the expected achievements.

7.4 COMPONENTS OF TIME SERIES

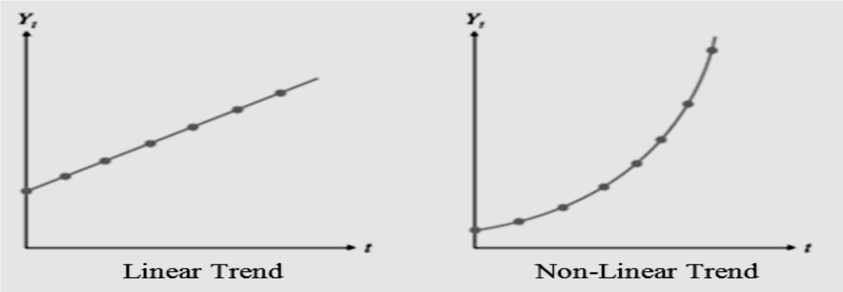
A large number of forces are there that affect the data. For example, if the sales of a company are changing with the passage of time, there are many forces responsible for it. We can classify these forces basically into four categories known as components or elements of the time series. Following are these components:

a) SECULAR TREND

The word is secular is taken from Latin word 'Saeculum', which means a ‘Generation’. So, as the names suggests, secular Trends are long-term Trends which normally occurs over it period of 15 to 20 years. Sometime, these trends may show upward results and other time it may show downward results. For example, we can see that number of persons who are travelling by air is increasing over a period of time. Similarly, we can see that infant mortality rate in the country is decreasing over a period of time. These both are secular trend but one trend is showing upward result and other trend is showing downward result.



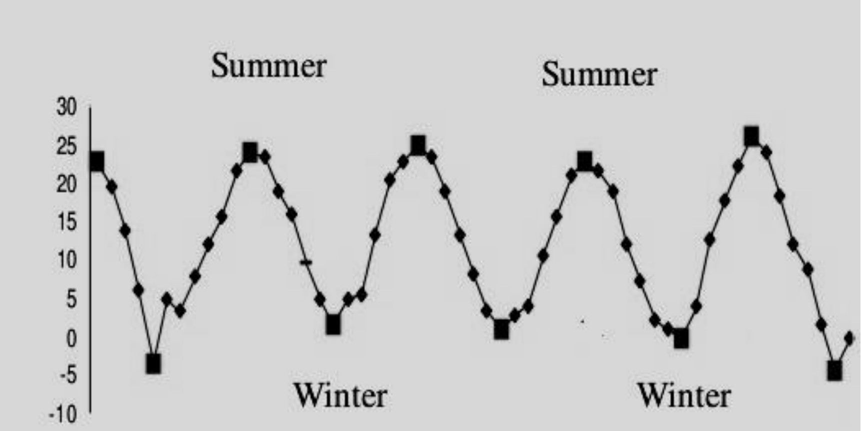
Upward Trend | Downward Trend | Stable Trend



b) SEASONAL VARIATIONS:

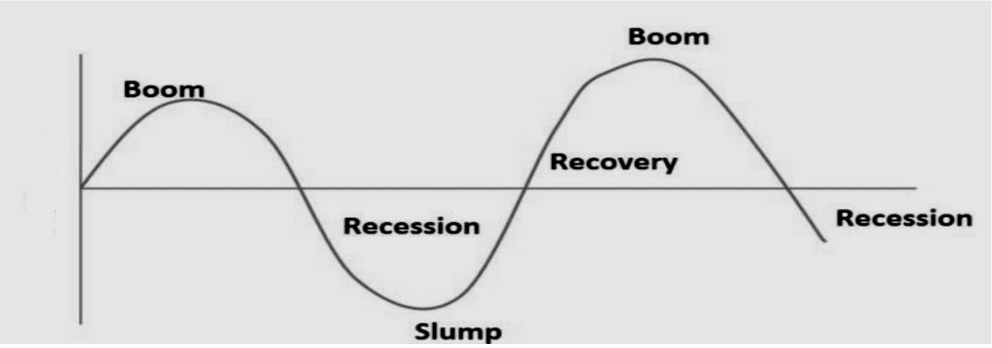
Seasonal variations are short term variations. These variations occur regularly and their trend is repetitive. These variations may occur every year, half yearly basis, monthly basis, weekly basis or any other time period basis. There may be many reasons for these variations but these variations generally occur due to following two reasons:

1. **Climatic conditions**: Sometimes seasonal variations take place due to climate change. We can see that there is climatic cycle that occur during the year. This climatic cycle also effects many things like sales of company, consumption patterns etc. For example, in rainy season sales of umbrellas increase, in summer season sales of air conditioners increase and similarly during the winter season sale of woollen clothes increases. These variations take place every year.
2. **Customs and traditions**: Sometimes seasonal variations take place due to customs and traditions. For example, in India is a tradition of purchasing new items in the household at the time of Diwali festival. So, this is also seasonal variation that take place every year.



c) CYCLICAL VARIATIONS

As the term 'cycle' suggest, these variations are recurrent variations. These variations are long Run variations and show a recurring pattern of rise and decline. These variations are also known as oscillating movements. These variations do not have any fixed duration. Sometime one cycle may be complete in 2-3 years, but some other times it may takes 7-8 years to complete. For example, a business cycle is cyclical variation that has four phases Boom, recession, depression and recovery.



d) IRREGULAR VARIATIONS

From the names of these variations, it is clear that these variations do not have any definite pattern and are irregular. These variations do not have any fixed time period and occur due to accidental or random factors like strikes, floods, pandemic, wars, earthquakes etc.

7.5 DECOMPOSITION OF TREND

As we have discussed above any time series data comprise various components namely Secular trend, seasonal variations, cyclical variations or irregular variations. In the time series Analysis, we try to identify various components of time series separately. This can be done by measuring the impact of one component while we keep another component constant. This process of finding each of the elements of time series separately is known as De-composition of time series. There are many models which are normally used to analyse the time series. These are:

7.5.1 Additive Model:

This model of decomposition assumes that the four elements of time series are not dependent on each other and does not affect each other. Each trend operates independently. So, if we have to measure overall trend of the time series, it is combination of all the four elements. By adding effect of all the elements, we can get the overall time series trend. Mathematically we can say that

$$\boxed{Y=T+S+C+I}$$

$$\boxed{\textbf{Short term fluctuations} = Y - T= S+C+I}$$

$$\boxed{\textbf{Cyclical and Irregular Fluctuation} = Y - T - S = C + I}$$

$$\boxed{\textbf{Irregular Fluctuation} = Y - T - S - C = I}$$

Where, Y = time series value, T = Secular Trend Variations, S = Seasonal Variations,
C = Cyclical Variations and I = Irregular Variations.

Though in additive model we assume that all the elements operate independently, but in reality, it is not true as all the elements have significant effects on each other and this is the major limitation of additive model.

7.5.2 Multiplicative Model:

The Multiplicative model is based on the assumption that all the components of the time series are related to each other and have significant effect on each other. So, if we want to calculate overall trend, it cannot be calculated by simply adding the four components. Rather it is multiple effect of all the four elements. So according to this model, overall trend is

$$\boxed{Y = T \times S \times C \times I}$$

$$\boxed{\textbf{Short term fluctuations} = \frac{Y}{T} = S \times C \times I}$$

$$\boxed{\textbf{Cyclical and Irregular Fluctuation} = \frac{Y}{T \times S} = C \times I}$$

$$\boxed{\textbf{Irregular Fluctuation} = = \frac{Y}{T \times S \times I} = I}$$

Here it is important to mention that the values of S, C and I are not absolute values rather these are relative variations and these are expressed in relative change or some indices.

7.6 MEASUREMENT OF TREND

Trend means the direction or tendency of series of data over a long period of time. We want to know whether the values are increasing over some time, decreasing over a period of time or these are stable over a period of time. This is known as Trend. We generally assume that the past behaviour of the data will continue in the future as well, so finding the trend could help us in predicting the future. There are four methods of finding the trend which are as follows:

- Free-hand graphic method
- Semi-average method
- Moving average method
- Method of least square

7.6.1 FREE HAND GRAPHIC METHOD

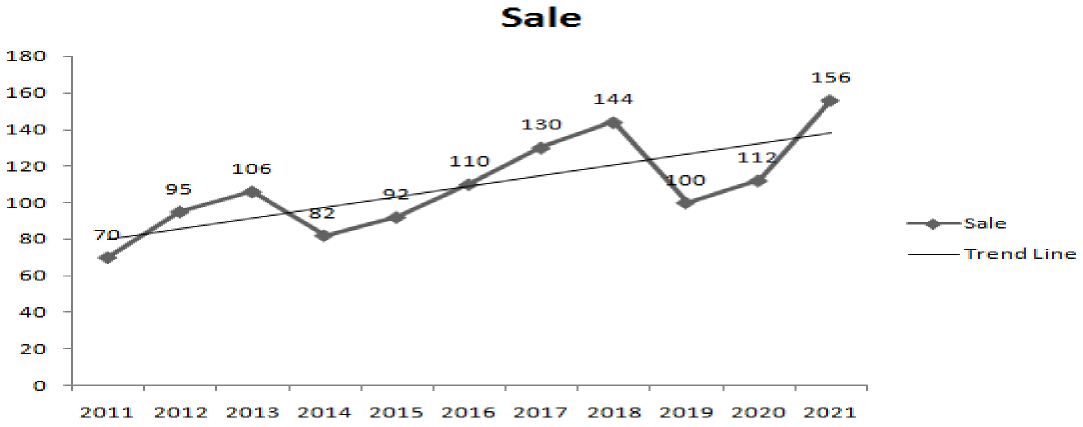
This is the simplest method of finding the trend and is very flexible. This method is also known as ‘free hand curve fitting method’. Following are the steps of finding trend under this method:

1. In the graph paper line chart is to be drawn.
2. For this purpose, time is taken on x-axis whereas values are taken on y axis.
3. Plot all the given values in the graph paper.
4. Then we join all the points in the graph paper to show the actual value.
5. After that smooth straight line is drawn which pass through the middle of the actual values drawn.
6. This line is the trend line.

Example 1: Fit the straight-line graphic curve from the following data:

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Sale	**70**	**95**	**106**	**82**	**92**	**110**	**130**	**144**	**100**	**112**	**156**

Solution:



From the above graph we can predict any value with the help of trend line.

CHECK YOUR UNDERSTANDING (A)

1. Following is the data of Harshit Ltd. draw a straight trend line using free hand graphic method.

Year:	2009	2010	2011	2012	2013	2014	2015	2016	2017

Sales (in '000 kg):	20	22	24	21	23	25	23	26	24

2. On basis of following data fit straight trend lines using free hand graphic method.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Production	64	82	97	71	78	112	115	131	88	100	146

7.6.2 SEMI AVERAGE METHOD

Semi average method is second method of finding the trend line. This is an objective method and is not merely based on guesswork. Under this method it is very easy to find trend line. Following are the steps of semi average method:

1. Divide the series into two equal parts, for example if there are 10 values take 5 values in each part.
2. In case of number of values are in odd number, middle value may be left and remaining values can be divided into two parts. For example, if there are 11 values, 6th value may be left and will have two parts having five values each.
3. Find the Arithmetic mean of both the parts.
4. These arithmetic means are called semi averages.
5. Now these semi averages are plotted in the graph as points against middle of each time period for which these have been calculated.
6. Join the points to find out straight line Trends.

Even Number of Years

Example 3: From the data given below find semi average trend line and also find out trend values.

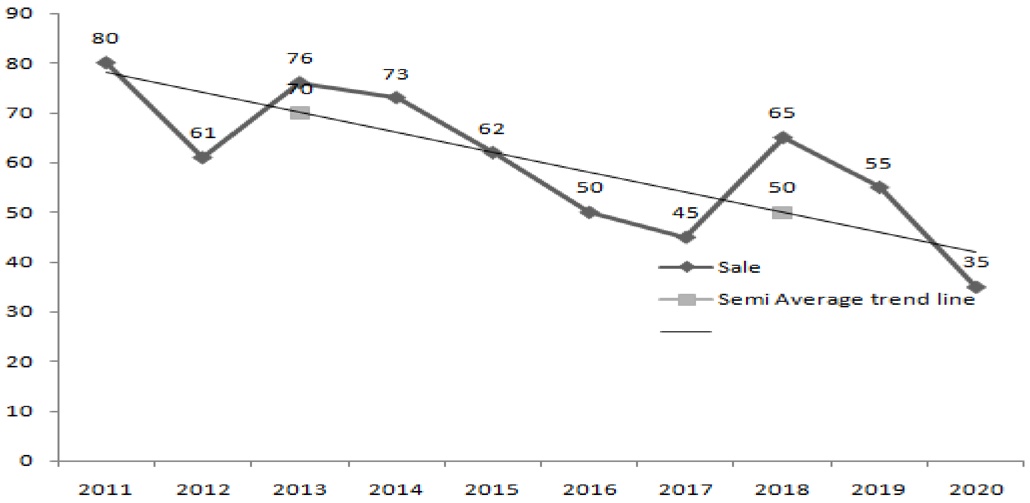
Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sale '000'	**80**	**61**	**76**	**73**	**62**	**50**	**45**	**65**	**55**	**35**

Solution: As the number of years is even, we have got two blocks of five years each. Now we will find arithmetic mean of these two blocks and will write against middle of the block.

Year	Sale ('000')	Semi Average
2011	80	
2012	61	
2013	76	$= \frac{80+61+76+73+62}{5} \quad = \frac{350}{5} = 70$

2014	73	
2015	62	
2016	50	
2017	45	
2018	65	$= \frac{50+45+65+55+35}{5} = \frac{250}{5} = 50$
2019	55	
2020	35	

For finding the trend in the graph 70 is plotted against year 2013 and 50 is plotted against the year 2018.



$$\text{Annual increment} = \frac{\textit{Difference in Semi Average values}}{\textit{Difference in two years to which Semi Average belongs}}$$

$$\text{Annual increment} = \frac{50-70}{2018-2013} = \frac{-20}{5} = -4$$

As we can see from the above data that semi average is showing downward trend so this annual increment will be deducted to semi average of 2013 onwards. For finding the values of the years before 2013 it will be added to the value every year. So, trend values are:

Year	Actual Sale ‘000’	Trend Sale ‘000’
2011	80	78 (74+4)
2012	61	74 (70+4)

2013	76	70
2014	73	66 (70-4)
2015	62	62 (66-4)
2016	50	58 (62-4)
2017	45	54 (58-4)
2018	65	50 (54-4)
2019	55	46 (50-4)
2020	35	42 (46-4)

Odd Number of Years

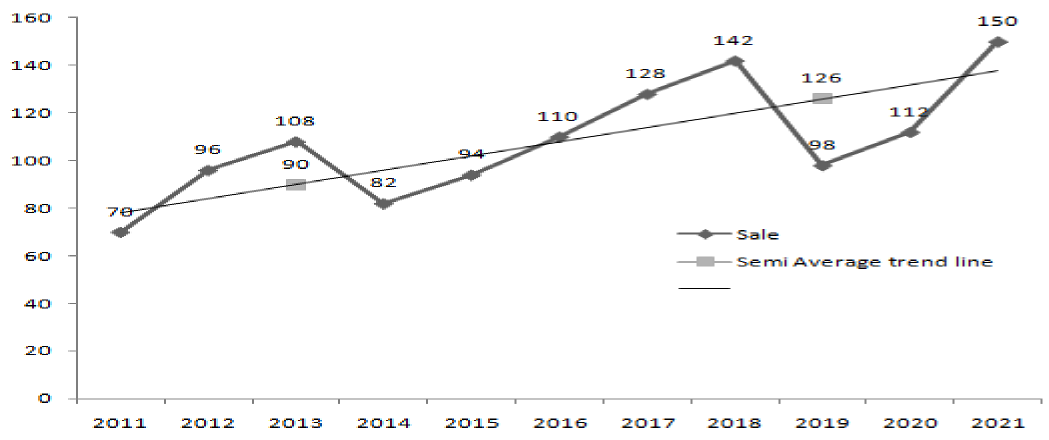
Example 4: From the data given below find semi average trend line and also find out trend values.

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Sale '000'	**70**	**96**	**108**	**82**	**94**	**110**	**128**	**142**	**98**	**112**	**150**

Solution: As the number of years is odd, the middle year 2016 is left and we have got two blocks of five years each. Now we will find arithmetic mean of these two blocks and will write against middle of the block.

Year	Sale '000'	Semi Average
2011	70	
2012	96	
2013	108	$= \frac{70+96+108+82+94}{5} = \frac{450}{5} = 90$
2014	82	
2015	94	
2016	110	
2017	128	
2018	142	
2019	98	$= \frac{128+142+98+112+150}{5} = \frac{630}{5} = 126$
2020	112	
2021	150	

For finding the trend in the graph 90 is plotted against year 2013 and 126 is plotted against the year 2019.



$$\text{Annual increment} = \frac{\textit{Difference in Semi Average values}}{\textit{Difference in two years to which Semi Average belongs}}$$

$$\text{Annual increment} = \frac{126 - 90}{2019 - 2013} = \frac{36}{6} = 6$$

As we can see from the above data that semi average is showing upward trend so this annual increment will be added to semi average of 2013 onwards. For finding the values of the years before 2013 it will be deducted from value every year. So, trend values are:

Year	Actual Sale '000'	Trend Sale '000'
2011	70	78 (84-6)
2012	96	84 (90-6)
2013	108	90
2014	82	96 (90+6)
2015	94	102 (96+6)
2016	110	108 (102+6)
2017	128	114 (108+6)
2018	142	120 (114+6)
2019	98	126 (120+6)
2020	112	132 (126+6)
2021	150	138 (132+6)

CHECK YOUR UNDERSTANDING (B)

1. From the production of Mahanta Ltd fits a straight-line trend using semi average method:

Year	2011	2012	2013	2014	2015	2016	2017	2018
Production ('000 Units)	200	210	218	192	204	216	224	228

Also predict value of 2020.

2. Fit straight line trend using Semi Average Method

Year	2012	2013	2014	2015	2016	2017	2018
Sales (in thousand units)	101	106	114	110	109	115	112

3. Sales of Abhinav are given, fit a straight-line trend using semi average method:

Year	2012	2013	2014	2015	2016	2017	2018
Sales (‘000 Units)	80	90	92	83	94	99	92

Also predict value of 2021.

Answers

1. 230, 3. 84

7.6.3 MOVING AVERAGE METHOD

Under this method we try to find out trend line using the concept of moving average. For this, first of all we decide the period for which moving average is to be calculated, for example, we can take 3 year moving average, 4 years moving average, 5 year moving average or so on. Following are the steps in this method

1. First of all, decide the length of period for which moving average will be taken.
2. Calculate the moving average of first group starting with first item.
3. After that find out moving average of second group leaving the first item.
4. Repeat this process until moving average is calculated for all the groups ending with last item.
5. Write the first moving average in front of the middle item of the group.
6. Repeat this process till all the moving averages are placed front of middle item of the group.
7. In case, even number of years are taken as period of moving average, the moving average is placed in middle of the period and then average of the adjacent averages is placed against mid item.

Odd period Moving Average

Example 5: Calculate 3 yearly and 5 yearly moving averages for the following data:

Year	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales	**52**	**49**	**55**	**49**	**52**	**57**	**54**	**58**	**59**	**60**	**52**	**48**

Solution: Following are the steps for 3 yearly moving average

1. First, compute the total value of first three years (2009, 2010, 2011) and place the three-year total against the middle year 2010.
2. Now, leaving the first year’s value, add up the values of the next three years (2010, 2011, 2012) and place the three-year total against the middle year 2011.

3. Repeat the process till last year's value i.e., 2020 is taken up.
4. Now divide the three year's total by 3 to get the average and place it in the next column. All these values represent the required trend values for the given year.
5. Same process can be repeated for 5 yearly moving average.

Year	Sale	3 Year Moving Total	3 Year Moving Average	5 Year Moving Total	5 Year Moving Average
2009	52				
2010	49	156	52		
2011	55	153	51	257	51.4
2012	49	156	52	262	52.4
2013	52	158	52.7	267	53.4
2014	57	163	54.1	270	54
2015	54	169	56.3	280	56
2016	58	171	57	288	57.6
2017	59	177	59	283	56.6
2018	60	171	57	277	55.4
2019	52	160	53.3		
2020	48				

Even period Moving Average:

Example 6: Calculate 4 yearly moving averages for the following data:

Year	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales	**250**	**260**	**275**	**300**	**290**	**310**	**318**	**325**	**350**	**340**

Solution: Following are the steps for 4 yearly moving average

1. First, compute the total value of first four years (2011, 2012, 2013, 2014) and place the four-year total in between 2nd and 3rd year i.e., between 2012 and 2013.
2. Now, leaving the first year's value, add up the values of the next four years (2012, 2013, 2014, 2015) and place the total b 2011 between 2013 and 2014.
3. Repeat the process till last year's value i.e., 2020 is taken up.
4. Now divide the four year's total by 4 to get the average and place it in the next column. All these values represent the required trend values for the given year.
5. Divide the first two four yearly average by 2 to get the required trend values corresponding to the given years as shown in the table:

Year	Value	4 Yearly Total	4 Yearly Average	Trend Value
2011	250			
2012	260			

		1085	271.25	
2013	275			276.25
		1125	281.25	
2014	300			287.5
		1175	293.75	
2015	290			299.12
		1218	304.5	
2016	310			307.63
		1243	310.75	
2017	318			318.25
		1303	325.75	
2018	325			329.5
		1333	333.25	
2019	350			
2020	340			

CHECK YOUR UNDERSTANDING (C)

1. Calculate 3 yearly moving averages for the following data:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Sales	11200	12300	10600	13400	13800	14500	11600	14300	13600	15400

2 Calculate 5 yearly moving averages for the following data:

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
No. of Employees	332	317	357	392	402	405	410	427	405	438

3 Calculate 4 yearly moving averages for the following data:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Value	100	105	115	90	95	85	80	65	75	70	75	80

Answers

1. 11366.7, 12100, 12600, 13900, 13300, 13466.7, 13166.7, 14433.3

2. 360, 374.6, 393.2, 407.2, 409.8, 417

3. 101.875, 88.75, 91.875, 84.375, 78.75, 74.375, 71.875, 73.125

7.6.4 LEAST SQUARE METHOD

This is most scientific and popular method of finding the trend line. Under this method, the lines of best fit are drawn as the lines trend. These lines are known as the lines of the best fit. According to the Least Square method, trend line should be plotted in such a way that sum of squares of the difference between actual value and estimated value of the dependent variable should be least or minimum possible. Mathematically this line is represented by

$$Yc = a + bX$$

Where Yc – Computed Trend Value

X – Independent Variable represented by time

a & b – Constants

Direct Methods to estimate Trend Line

Following are steps for finding trend line with help of Direct Method:

1. Take the problem with two variables with X variable as time and other variable for which trend is to be computed like sales, population etc. represented by Y.
2. Assume first year as base year and put the value ‘0’ against it.
3. Now put value 1 against second year, 2 against third year and so on till all the years are covered.
4. Now find the values of $\sum X$, $\sum X^2$, $\sum XY$ from the given values.
5. Put these values in following equation:

   $$\Sigma Y = na + b\Sigma X$$

   $$\Sigma XY = a\,\Sigma X + b\Sigma X^2$$
6. Solve these equations simultaneously and find the values of ‘a’ and ‘b’.
7. Put value of ‘a’ and ‘b’ in trend equation Yc = a + bX.
8. Now this trend equation can be used for finding the trend values.

Example 7. The data of sales of Alpha Ltd is given for last 9 years. Based on the data find trend value of the year 2021 using the method of least square.

Year	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales ‘000’	**10**	**12**	**15**	**20**	**30**	**40**	**50**	**60**	**70**

Solution:

Year	X	Sales (Y)	X^2	XY
2012	0	10	0	0
2013	1	12	1	12
2014	2	15	4	30
2015	3	20	9	60
2016	4	30	16	120
2017	5	40	25	200
2018	6	50	36	300
2019	7	60	47	420
2020	8	70	64	560
	$\sum X = 36$	$\sum Y = 307$	$\sum X^2 = 204$	$\sum XY = 1702$

This is given by Y = *a* + *bX*

where *a* and *b* are the two constants that are found by solving simultaneously the two normal equations as follows:

$\Sigma Y = na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the given values in the above equations we get,

$307 = 9a + 36b$ …………………………………….. (i)

$1702 = 36a + 204b$ …………………………………….. (ii)

Multiplying the eqn. (i) by 4 we get

$1228 = 36a + 144b$…………………………………….. (iii)

Subtracting the equation (iii) from equation (ii) we get,

$1702 = 36a + 204b$

$\underline{-1228 = -36a - 144b}$

$474 = 60b$

or $b = 7.9$

3 $\Sigma Y = na + b\Sigma X$

If $\Sigma X = 0$ than $\Sigma Y = na$

$a = \frac{\Sigma Y}{n}$

Equation (ii) $\Sigma XY = a\Sigma X + b\Sigma X^2$

If $\Sigma X = 0$ than $\Sigma XY = b\Sigma X^2$

$b = \frac{\Sigma XY}{\sum X^2}$

Odd number of Years

Example 8. The data of sales of Mahesh and Co is given for last 7 years. On the basis of the data find trend line using the method of least square and find trend value of 2021.

Year	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales '000'	**672**	**824**	**967**	**1204**	**1464**	**1758**	**2057**

Solution: Since the number of years is odd, 2017 is taken as base year with a value of 0 and one year is taken as one unit.

Year	X	Sales (Y)	X^2	XY
2014	-3	672	9	-2016
2015	-2	824	4	-1648
2016	-1	967	1	-967
2017	0	1204	0	0
2018	1	1464	1	1464
2019	2	1758	4	3516
2020	3	2057	9	6171
	$\sum X = 0$	$\sum Y = 8946$	$\sum X^2 = 28$	$\sum XY = 6520$

As ∑X is 0, we can apply short cut method

$$a = \frac{\Sigma Y}{n} = \frac{8946}{7} = 1278$$

$$b = \frac{\Sigma XY}{\sum X^2} = \frac{6520}{28} = 232.9$$

Putting these values in the equation Y = a + bX we get

Y = 1278+ 232.9 X

So, if we want to calculate the trend value of the year 2021 the value of X will be 4 (as 2017 is our base year and its value is 0), the value of Y will be

Y = 1278+ 232.9 (4) = 2209.6

Odd number of Years

Example 9. The data of sales of Abhilasha Ltd is given for last 8 years. On the basis of the data find trend line using the method of least square and find trend value of 2021.

Year	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Sales '000'	**80**	**90**	**92**	**83**	**94**	**99**	**92**	**104**

Solution: Since the number of years is even, so will take the origin as midpoint of 2016 and 2017 and further for the sake of simplicity one year is taken as two units (6 Months as 1 unit).

Year	X	Sales (Y)	X^2	XY
2013	-7	80	49	-560
2014	-5	90	25	-450
2015	-3	92	9	-276

2016	-1	83	1	-83
2017	1	94	1	94
2018	3	99	9	297
2019	5	92	25	460
2020	7	104	49	728
	$\sum X = 0$	$\sum Y = 734$	$\sum X^2 = 168$	$\sum XY = 210$

As ∑X is 0, we can apply short cut method

$$a = \frac{\Sigma Y}{n} = \frac{7346}{8} = 91.75$$

$$b = \frac{\Sigma XY}{\sum X^2} = \frac{210}{168} = 1.25$$

Putting these values in the equation Y = *a* + bX we get

Y = 91.75+ 1.25 X

So, if we want to calculate the trend value of the year 2021 the value of X will be 9 (as mid of 2016 and 2017 is our base year and 1 year is taken as 2 units), the value of Y will be

Y = 91.75+ 1.25 (9) = 103

CHECK YOUR UNDERSTANDING (D)

1. These are the number of salesmen working in Alpha Ltd:

Year	2011	2012	2013	2014	2015	2016
Salesmen	28	38	46	40	56	60

Fit straight-line trend using the method of least squares.

2 Fit a straight-line trend by Method of least square and estimate the exports of 2021 using the short cut method:

Year	2013	2014	2015	2016	2017	2018	2019	2020
Exports	15	20	24	29	35	45	60	85

3 Determine the equation of straight line which best fits the following data

Year	2012	2013	2014	2015	2016	2017	2018	2019	2020
Value	620	713	833	835	810	745	726	806	861

4 Determine the equation of straight line which best fits the following data

Year	2001	2002	2004	2006	2007
Sales ‘Lacs’	5	8	12	20	25

5 Determine the equation using method of least square from a number of accidents from the following data and find trend values also.

Year	2001	2002	2003	2004	2005	2006	2007	2008
Accidents	38	40	65	72	69	60	87	95

Answers

1. Y = 44.67 + 2.97 X,
2. Y = 39.125 + 4.517 X; value of 2021 – 79.778
3. Y = 709.51 + 15.65 X
4. Y = 14 + 3.23 X (taking 2004 as year of origin.
5. Y = 65.75 + 3.667 X (taking 2004.5 as year of origin)

Trend Values 40.081, 47.415, 54.749, 62.083, 69.417, 76.751, 84.085, 91.419

7.7 MEANING OF INDEX NUMBER

An index number is a statistical tool that measures the changes in the data over the period. Index number is not a new tool used in statistics, rather the use of index numbers is very old. As per available records, index number was first time constructed in the year 1764 by an Italian named Carli. In his index number, Carli compared the prices of the Year 1750 with the price level of the year 1500. Though normally index numbers are used for measuring the change in price over a period of time, but hardly there is any area in Economics or Commerce where Index numbers are not used. Different types of index numbers are used in economics such as Industrial Production Index, Agricultural Production Index and Population Index etc.

According to Croxton and Cowden, "Index numbers are devices for measuring differences in the magnitude of a group of related variables."

7.8 USES OF INDEX NUMBERS

1. Index number is a very powerful tool for economic and business analysis. We often call the index number ‘Barometer of the Economy’. With the help of Index Number, we can see pulse of the economy.
2. Index number is a very helpful tool in planning activities and formulation of business policy.
3. With the help of index numbers, economists try to find out trends in prices, production, imports and exports, etc.
4. Index number shows the cost of living over a period of time. This also helps government in fixing the wage rate of the labour.
5. Index number also helps us in calculation of Real National Income of the country.

7.9 LIMITATIONS OF INDEX NUMBERS

1. As index numbers are based on sample data, these can give only approximate results not the accurate result.
2. Index number normally deals with one variable, so it is not possible to calculate a single index for all economic activities.
3. There is not a single standard method of calculating index. Different experts calculate index in their way.
4. Index number are special types of average, it does not deal with all the situations.
5. Finding the appropriate base period is very difficult in construction of index number.

7.10 PROBLEMS IN CONSTRUCTION OF INDEX NUMBERS

- Purpose of Index Numbers
- Selection Of Base Year
- Selection of Number of Items or Commodities
- Selection of Source of Data
- Selection of the Average
- Selection of Appropriate Weight
- Selection of appropriate formula

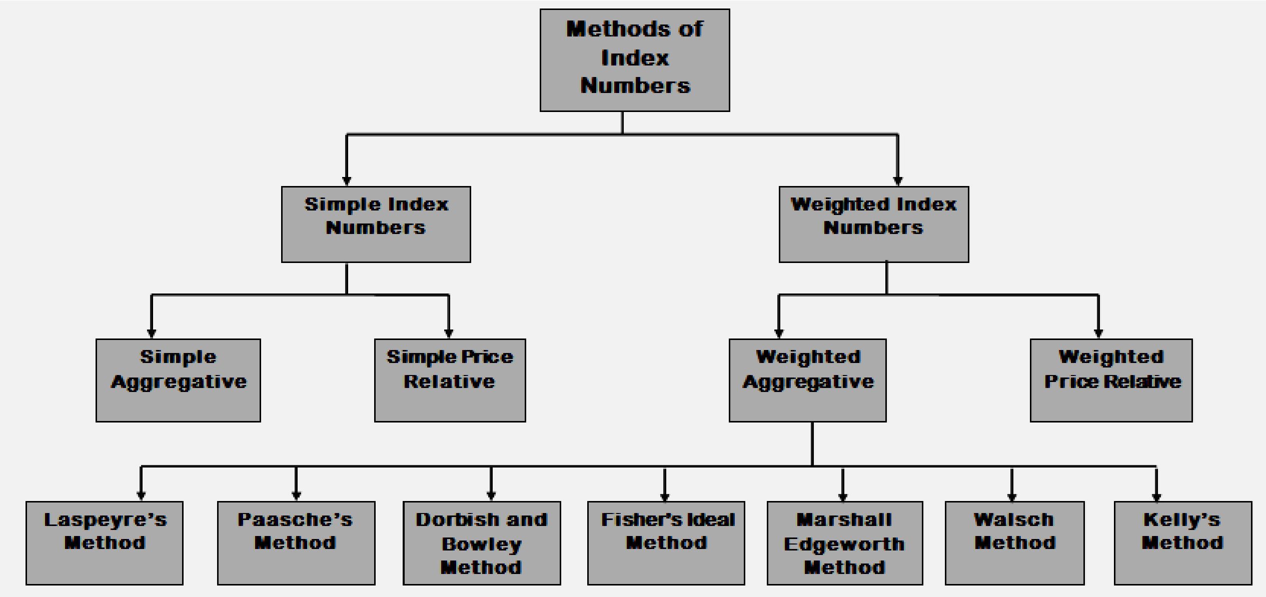
7.11 DIFFERENT TYPES OF INDEX NUMBERS

1. **Price Index Numbers**: These index numbers are used for measuring the change in prices of commodities over a period of time. In other words, we can say that these index numbers find the change in value of money over a period of time. These index numbers are most popular. These Index numbers may be based on Wholesale Price Index or Retail Price Index.
2. **Quantity Index Numbers**: The Quantity or Volume Index Numbers measure the change in quantities used by people over a period of time. under these index numbers, we calculate change in physical quantity of goods produced, consumed or sold over a period of time. There are different types of quantity index numbers such as Agricultural Production Index Number, Industrial Production Index Number, Export Import Index Number etc.
3. **Value Index Numbers**: Value Index Numbers compare the change in total value over a period. These index numbers take into consideration both prices and quantity of the product

while finding the change over a period of time. These Index Numbers are very useful in finding consumption habits of the consumers.

7.12 DIFFERENT METHODS OF INDEX NUMBERS

As we have already discussed, an Index number is a device that shows changes in price over a period of time. Now a question arises about how to calculate the index number. There are a number of methods for preparing the index numbers. The following chart shows various methods of preparing index numbers.



7.12.1 SIMPLE INDEX NUMBER

This further divided into the simple aggregative and simple price relative

7.12.1.1 SIMPLE AGGREGATIVE METHOD

This is one of the old and simple methods of finding the index number. Under this method we calculate the index number of a given period by dividing the aggregate of all the prices of the current year by the aggregate of all the prices of the base year. After that we multiply the resultant figure with 100 to find the index number. Following are the steps:

1. Decide the base year.
2. Add all the prices of base year for all available commodities, it is denoted by $\sum P_0$.
3. Add all the prices of base year for all available commodities, it is denoted by $\sum P_1$.
4. Use following the formula for calculating index number under this method:

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where, P01 – Price Index Number of Current Year

$\sum P_1$ – Aggregate of Prices of Current Year

$\sum P_0$ – Aggregate of Prices of Base Year

Example 1. Construct Simple Aggregative Index number of the year 2020by taking the base as prices of 2015.

Commodity	**Price of the Year 2015**	**Price of the Year 2020**
Wheat	**20**	**26**
Sugar	**40**	**34**
Oil	**60**	**120**
Pulses	**80**	**140**

Solution: Price Index (Year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1
Wheat	20	26
Sugar	40	34
Oil	60	120
Pulses	80	140
	$\sum P_0 = 200$	$\sum P_1 = 320$

$$\text{Price Index } (P_{01}) = \frac{\sum P_1}{\sum P_0} \times 100 = \frac{320}{200} \times 100 = 160$$

The price index shows that prices have increased by 60% in 2020 than 2015.

7.12.1.2 SIMPLE PRICE RELATIVE METHOD

This method is a bit improvement over the simple aggregative method. Simple aggregative method is affected by the magnitude of the price of the item. However, this method is not affected by magnitude of the price of item. Further, in this method it is not necessary to use Arithmetic mean as average rather we can use any method of finding average, such as Arithmetic mean, Geometric mean, Median, Mode etc. However, normally we prefer to use Arithmetic mean in this case. Following are the steps of this method:

1. Decide the base year.
2. Calculate the price relative to current year for each commodity by dividing current Prices (P_1) with base year price (P_0) using the following formula $\frac{P_1}{P_0}. \times 100$
3. Find sum of all the price relatives so calculated.

4. Divide the sum or price relatives by number of items to get index number by using the following formula:

$$P_{01} = \frac{\sum\frac{P_1}{P_0} \times 100}{N}$$

Example 2. Construct Simple Price Relative Index number of the year 2020by taking the base as prices of 2015.

Commodity	**Price of the Year 2015**	**Price of the Year 2020**
Wheat	**20**	**26**
Sugar	**40**	**34**
Oil	**60**	**120**
Pulses	**80**	**140**

Solution: Price Index (Year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1	Price Relative $\frac{P_1}{P_0}$. × 100
Wheat	20	26	$\frac{26}{20}$. × 100 = 130
Sugar	40	34	$\frac{34}{40}$. × 100 = 85
Oil	60	120	$\frac{120}{60}$. × 100 = 200
Pulses	80	140	$\frac{140}{80}$. × 100 = 175
			$\sum\frac{P_1}{P_0}$. × 100 = 590

Price Index (P_{01}) $= \frac{\sum\frac{P_1}{P_0}\times 100}{N} = \frac{590}{4} = 147.50$

Price index shows that prices have increased by 47.5% in 2020 than 2015.

CHECK YOUR PROGRESS (D)

1. Calculate Index number for 2015 taking 209 as base using Simple Aggregative Method and Simple Average of Relatives Method:

Items	Price 2011	Price 2015
A	350	510
B	45	40
C	77	156
D	37	47
E	10	12

2. Find index using simple average of price relative using 2017 as base.

Items	Price 2017	Price 2019
A	15	30

B	18	24
C	16	20
D	14	21
E	25	35
F	40	30

3. Find simple aggregative index

Items	P_0	P_1
Oil	60	70
Pulses	70	60
Rice	50	40
Sugar	40	40

Answers

1. 147.4, 132.84, 2. 137.22, 3. 95.45

7.12.2 WEIGHTED INDEX NUMBER

Which is further divided into weighted aggregative and weighted price relative method

7.12.2.1 WEIGHTED AGGREGATIVE PRICE INDEX

Simple Aggregative methods of Index Numbers assume that all the items of Index Number are equally important. There is no item which is more important than other. So, this method provides equal weightage to all items. However, in practical life it is not true. Some items carry more importance than other items, for example in human life expenditure on food carries more importance than expenditure on entertainment. So, we have weighted method of index numbers which considers relative importance of the item also.

Weighted Aggregative Method is one such method. This method is more or less same as the Simple Aggregative Method but main difference is that it also considers relative weights of the items. Generally, the quantity of the item consumed is considered as weight in this case. There are many methods of calculating Weighted Aggregative Price Index which are discussed as follows:

a) Laspeyre's Method:

This method was suggested by Mr. Laspeyre in 1871. Under this method base year quantities of the various products are assumed as weight for preparing the index numbers. The following steps may be used:

1. Multiply Prices of the base year (P_0) with the quantities of the base year (Q_0) for every commodity.
2. Add the values calculated in step 1, the sum is denoted as $\sum P_0 Q_0$

3. Multiply Prices of the current year (P_1) with the quantities of the base year (Q_0) for every commodity.
4. Add the values calculated in step 3, the sum is denoted as $\sum P_1 Q_0$.
5. Use following formula for calculating index number:

$$P_{01} = \frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times 100$$

b) Paasche's Method:

This method was suggested by Mr. Paasche in 1874. Under this method current year quantities of the various products are assumed as weight for preparing the index numbers. The following steps may be used:

1. Multiply Prices of the base year (P_0) with the quantities of the current year (Q_1) for every commodity.
2. Add the values calculated in step 1, the sum is denoted as $\sum P_0 Q_1$
3. Multiply Prices of the current year (P_1) with the quantities of the current year (Q_1) for every commodity.
4. Add the values calculated in step 3, the sum is denoted as $\sum P_1 Q_1$.
5. Use following formula for calculating index number:

$$P_{01} = \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100$$

c) Dorbish and Bowley's Method:

This method is based on both Lasypeyre's Method and Paasche's Method, that's why this method is also known as L-P formula. Under this method we calculate the index number by taking the arithmetic mean of the formula given by Laspeyre and Paasche. So, following formula is used in case of the Dorbish and Bowley Method:

$$P_{01} = \frac{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} + \frac{\sum P_1 Q_1}{\sum P_0 Q_1}}{2} \times 100$$

d) Fisher's Ideal Index Method:

This method was suggested by Prof Irving Fisher and it is assumed as one of the best methods of constructing the Index Number. That's why this method is also called Ideal Index Number. This method is based on both Lasypeyre's Method and Paasche's Method, but instead of taking arithmetic mean of both formulas, Fisher used the geometric mean on the formula given by Laspeyre and Paasche. So, following formula for calculating Fisher's ideal Index number:

$$\sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times 100$$

Fisher's Method is called ideal index number for to following reasons:

1. This method uses geometric mean as base which is perhaps best average for constructing Index numbers.
2. This method considers both quantities of base year as well as current year as weight.
3. This method satisfies both the time reversal and factor reversal tests.
4. It is comprehensive method and cover all values of data i.e P_0, Q_0, P_1, Q_1 etc.

e) Marshal Edgeworth Index Method:

Like Fisher's method, this method also uses the quantities of base as well as current year as weight. Under this method arithmetic mean of the quantity of base and current year is assumed as weight. This method is comparatively simple than Fisher's method as it does not use complex concept of Geometric mean. Following is the formula of this method.

$$P_{01} = \frac{\sum P_1 (Q_0 + Q_1)}{\sum P_0 (Q_0 + Q_1)} \times 100 \text{ or}$$

$$\frac{\sum P_1 Q_0 + \sum P_1 Q_1}{\sum P_0 Q_0 + \sum P_0 Q_1} \times 100$$

Example 3. Construct Weighted Aggregative Index number of the year 2020 by taking the base as prices of 2015 using Laspeyre, Paasche, Dorbish & Bowley, Fisher, Marshal Edgeworth and Kelly's method.

Item	Price of the Year 2015	Quantity of the Year 2015	Price of the Year 2020	Quantity of the Year 2020
A	6	50	10	56
B	2	100	2	120
C	4	60	6	60
D	10	30	12	24
E	8	40	12	36

Solution:

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1

A	6	50	10	56	300	336	500	560
B	2	100	2	120	200	240	200	240
C	4	60	6	60	240	240	360	360
D	10	30	12	24	300	240	360	288
E	8	40	12	36	320	288	480	432
					$\sum P_0Q_0$ = 1360	**$\sum P_0Q_1$ = 1344**	**$\sum P_1Q_0$ = 1900**	**$\sum P_1Q_1$ = 1880**

1. Laspeyre's Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{1900}{1360} \times 100 = 139.71$$

2. Paasche's Method:

$$P_{01} \quad \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{1880}{1344} \times 100 = 139.88$$

3. Dorbish and Bowley's Method:

$$P_{01} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times 100$$

$$= \frac{\frac{1900}{1360} + \frac{1880}{1344}}{2} \times 100 = \frac{2.796}{2} = 139.79$$

4. Fisher's Ideal Index Method:

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times 100$$

$$= \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}} \times 100 = \sqrt{1.9543} \times 100 = 139.79$$

5. Marshal Edgeworth Index Method:

$$\frac{\sum P_1Q_0 + \sum P_1Q_1}{\sum P_0Q_0 + \sum P_0Q_1} \times 100$$

$$= \frac{1900 + 1880}{1360 + 1344} \times 100 = \frac{3780}{2704} \times 100 = 139\ .79$$

Example 4. Construct a Weighted Aggregative Index number using Laspeyre, Paasche, Dorbish & Bowley and Fisher, method.

Item	Price of the Base Year	Expenditure of the Base Year	Price of the Current Year	Expenditure of the Current Year
A	**2**	**40**	**5**	**75**
B	**4**	**16**	**8**	**40**
C	**1**	**10**	**2**	**24**

D	**5**	**25**	**10**	**60**

Solution: We know that Expenditure = Price × Quantity

$$\text{So, Quantity} = \frac{Expenditure}{Price}$$

Item	P₀	Q₀	P₁	Q₁	P₀Q₀	P₀Q₁	P₁Q₀	P₁Q₁
A	2	20	5	15	40	30	100	75
B	4	4	8	5	16	20	32	40
C	1	10	2	12	10	12	20	24
D	5	5	10	6	25	30	50	60
					$\sum P_0Q_0$ = 91	**$\sum P_0Q_1$ = 92**	**$\sum P_1Q_0$ = 202**	**$\sum P_1Q_1$ = 199**

1. Laspeyre’s Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{202}{91} \times 100 = 221.98$$

2. Paasche’s Method:

$$P_{01} \quad \frac{\sum P_1Q_1}{\sum P_0Q_1} \times 100 = \frac{199}{92} \times 100 = 216.39$$

3. Dorbish and Bowley’s Method:

$$P_{01} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times 100$$

$$= \frac{\frac{202}{91} + \frac{199}{92}}{2} \times 100 = \frac{4.3828}{2} = 219.14$$

4. Fisher’s Ideal Index Method:

$$\sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times 100$$

$$= \sqrt{\frac{202}{191} \times \frac{199}{92}} \times 100 = \sqrt{4.8015} \times 100 = 219.12$$

7.12.2.2 WEIGHTED PRICE RELATIVE METHOD

This method almost similar to simple price relative method. However, simple price relative gives equal importance to all items under consideration. But in our life, all items do not carry equal importance. Some items are more important or on some items we spend more amount. Changes in price of some items affect us more than changes in price of some other items. So, we have weighted price relative method. This method is similar to simple price relative method but also assigns weights to the items. Further, in this method it is not necessary to use Arithmetic mean as average rather we can use any method of finding average, such as Arithmetic mean, Geometric mean, etc. However, normally we prefer to use Arithmetic mean in this case. Following are the steps of this method:

1. Decide the base year.
2. Calculate the price relative of current year for each commodity by dividing current Prices (P_1) with base year price (P_0) using the following formula $\frac{P_1}{P_0}. \times 100$.
3. Find the weights of the items to be assigned.
4. Multiply price relative so calculated with the weights and find out the product of both.
5. Find sum of product so calculated.
6. Find sum of the weights assigned.
7. Divide the sum of the weighted price relatives by sum of weights to get index number by using the following formula:

$$P_{01} = \frac{\sum W \frac{P_1}{P_0} \times 100}{\sum W}$$

Example 5. Construct Weighted Price Relative Index number of the year 2020 by taking the base as prices of 2015.

Commodity	Price of the Year 2015	Price of the Year 2020	Weights
Wheat	20	26	40
Sugar	40	34	5
Oil	60	120	3
Pulses	80	140	2

Solution: Price Index (Year 2015 taken as the base year)

Commodity	Price of the Year 2015 P_0	Price of the Year 2020 P_1	Price Relative $\frac{P_1}{P_0}. \times 100$	Weights (W)	Weighted Price Relatives $W \frac{P_1}{P_0}. \times 100$
Wheat	20	26	$\frac{26}{20}. \times 100 = 130$	40	5200
Sugar	40	34	$\frac{34}{40}. \times 100 = 85$	5	424
Oil	60	120	$\frac{120}{60}. \times 100 = 200$	3	600
Pulses	80	140	$\frac{140}{80}. \times 100 = 175$	2	350
				$\sum W = 50$	$\sum W \frac{P_1}{P_0}. \times 100 = 6575$

Price Index (P_{01}) = $\frac{\sum W\frac{P_1}{P_0}\times 100}{\sum W} = \frac{6575}{50}$ = 131.50

Price index shows that prices have increased by 31.5% in 2020 than 2015.

7.13 TESTS OF CONSISTENCY FOR INDEX NUMBERS

There are a number of methods through which index numbers can be calculated. Each method has its own merits and demerits. Now a question is which of these methods can be treated best. In order to find out which method is better than others, there are four tests. If any index number satisfies these tests, we may consider the index number to be ideal one.

7.13.1 Unit Test

Unit test says that any index number can be treated as ideal only if it is free from the unit in which quantity is measured. Whether prices are quoted for single item or dozen items, the index number must not be affected by the same. Only simple average of price relative method satisfies this condition.

7.13.2 Time Reversal Test

This test was suggested by Fisher. According to this test, an ideal index number works both ways i.e., backward and forward. So, if index is prepared by taking old period as base year and new period as current year it comes to be 200 which means prices in current period are doubled. Now say reverse is done, new period is taken as base and old period are taken as current year, this test says that index should be 50 which means earlier prices were half of current prices. In other words, we can say that following conditions should be satisfied

$$\boxed{\mathbf{P_{01} \times P_{10} = 1}}$$

Following is the formula of time reversal test in different cases:

1. **Laspeyre's Method:**

$$P_{01} \times P_{10} = \frac{\Sigma P_1Q_0}{\Sigma P_0Q_0} \times \frac{\Sigma P_0Q_1}{\Sigma P_1Q_1} \neq 1$$

This method does not satisfy time reversal test.

2. **Paasche's Method:**

$$P_{01} \times P_{10} = \frac{\Sigma P_1Q_1}{\Sigma P_0Q_1} \times \frac{\Sigma P_0Q_0}{\Sigma P_1Q_0} \neq 1$$

This method does not satisfy time reversal test.

3. Dorbish and Bowley's Method:

$$P_{01} \times P_{10} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times \frac{\frac{\sum P_0Q_1}{\sum P_1Q_1} + \frac{\sum P_0Q_0}{\sum P_1Q_0}}{2} \neq 1$$

This method does not satisfy time reversal test.

4. Fisher's Ideal Index Method:

$$P_{01} \times P_{10} = \sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times \sqrt{\frac{\sum P_0Q_1}{\sum P_1Q_1} \times \frac{\sum P_0Q_0}{\sum P_1Q_0}} = 1$$

This method satisfies time reversal test.

5. Marshal Edgeworth Index Method:

$$P_{01} \times P_{10} = \frac{\sum P_1Q_0 + \sum P_1Q_1}{\sum P_0Q_0 + \sum P_0Q_1} \times \frac{\sum P_0Q_1 + \sum P_0Q_0}{\sum P_1Q_1 + \sum P_1Q_0} = 1$$

This method satisfies time reversal test.

7.13.3 Factor Reversal Test (F.R.T.)

This test was also suggested by Fisher. According to this test, an ideal index number does not give inconsistent results if we change price with quantity and quantity with price. According to this test when we multiply change in price by change in quantity the ratio must be equal to total change in value.

$$\boxed{\mathbf{P_{01} \times Q_{01}} = \frac{\sum P_1Q_1}{\sum P_0Q_0}}$$

Following is the formula of factor reversal test in different cases:

1. Laspeyre's Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum Q_1P_0}{\sum Q_0P_0} \neq \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method does not satisfy factor reversal test.

2. Paasche's Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1Q_1}{\sum P_0Q_1} \times \frac{\sum Q_1P_1}{\sum Q_0P_1} \neq \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method does not satisfy factor reversal test.

3. Dorbish and Bowley's Method:

$$P_{01} \times Q_{10} = \frac{\frac{\sum P_1Q_0}{\sum P_0Q_0} + \frac{\sum P_1Q_1}{\sum P_0Q_1}}{2} \times \frac{\frac{\sum Q_1P_0}{\sum Q_0P_0} + \frac{\sum Q_1P_1}{\sum Q_0P_1}}{2} \neq \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method does not satisfy factor reversal test.

4. Fisher's Ideal Index Method:

$$P_{01} \times Q_{10} = \sqrt{\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_1Q_1}{\sum P_0Q_1}} \times \sqrt{\frac{\sum Q_1P_0}{\sum Q_0P_0} \times \frac{\sum Q_1P_1}{\sum Q_0P_1}} = \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method satisfies factor reversal test.

5. Marshal Edgeworth Index Method:

$$P_{01} \times Q_{10} = \frac{\sum P_1Q_0+\sum P_1Q_1}{\sum P_0Q_0+\sum P_0Q_1} \times \frac{\sum Q_1P_0+\sum Q_1P_1}{\sum Q_0P_0+\sum Q_0P_1} = \frac{\sum P_1Q_1}{\sum P_0Q_0}$$

This method does not satisfy factor reversal test.

Example 6. Construct Weighted Aggregative Index numbers using Laspeyre, Paasche, and Fisher methods also check whether these satisfy T.R.T. and F.R.T or not

Item	Price of the Base Year	Qty. of the Base Year	Price of the Current Year	Qty. of the Current Year
A	**30**	**7**	**40**	**5**
B	**40**	**12**	**60**	**8**
C	**60**	**10**	**50**	**15**
D	**30**	**15**	**20**	**18**

Solution:

We know that Expenditure = Price × Quantity

So, Quantity = $\frac{Expenditure}{Price}$

Item	P_0	Q_0	P_1	Q_1	P_0Q_0	P_0Q_1	P_1Q_0	P_1Q_1
A	30	7	40	5	210	150	280	200
B	40	12	60	8	480	320	720	480
C	60	10	50	15	600	900	500	750
D	30	15	20	18	450	540	300	360
					$\sum P_0Q_0$ = 1740	**$\sum P_0Q_1$ = 1910**	**$\sum P_1Q_0$ = 1800**	**$\sum P_1Q_1$ = 1790**

1. Laspeyre's Method:

$$P_{01} = \frac{\sum P_1Q_0}{\sum P_0Q_0} \times 100 = \frac{1800}{1740} \times 100 = 103.45$$

Time Reversal Test

$$\frac{\sum P_1Q_0}{\sum P_0Q_0} \times \frac{\sum P_0Q_1}{\sum P_1Q_1} = \frac{1800}{1740} \times \frac{1910}{1790} \neq 1$$

It does not satisfy time reversal test.

Factor Reversal Test

$$\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum Q_1 P_0}{\sum Q_0 P_0} = \frac{1800}{1740} \times \frac{1910}{1740} \neq \frac{1790}{1740}$$

It does not satisfy facto reversal test.

2. **Paasche's Method:**

$$P_{01} \quad \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times 100 = \frac{1790}{1910} \times 100 = 93.72$$

Time Reversal Test

$$= \frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times \frac{\sum P_0 Q_0}{\sum P_1 Q_0} = \frac{1790}{1910} \times \frac{1740}{1800} \neq 1$$

It does not satisfy time reversal test.

Factor Reversal Test

$$\frac{\sum P_1 Q_1}{\sum P_0 Q_1} \times \frac{\sum Q_1 P_1}{\sum Q_0 P_1} = \frac{1790}{1910} \times \frac{1790}{1800} \neq \frac{1790}{1740}$$

It does not satisfy factor reversal test.

3. **Fisher's Ideal Index Method:**

$$\sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times 100$$

$$= \sqrt{\frac{1800}{1740} \times \frac{1790}{1910}} \times 100 = \sqrt{.96948} \times 100 = 98.462$$

Time Reversal Test

$$= \sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times \sqrt{\frac{\sum P_0 Q_1}{\sum P_1 Q_1} \times \frac{\sum P_0 Q_0}{\sum P_1 Q_0}} \neq 1$$

$$= \sqrt{\frac{1800}{1740} \times \frac{1790}{1910}} \times \sqrt{\frac{1910}{1790} \times \frac{1740}{1800}} = 1$$

It satisfies time reversal test.

Factor Reversal Test

$$\sqrt{\frac{\sum P_1 Q_0}{\sum P_0 Q_0} \times \frac{\sum P_1 Q_1}{\sum P_0 Q_1}} \times \sqrt{\frac{\sum Q_1 P_0}{\sum Q_0 P_0} \times \frac{\sum Q_1 P_1}{\sum Q_0 P_1}} = \frac{\sum P_1 Q_1}{\sum P_0 Q_0}$$

$$\sqrt{\frac{1800}{1740}\times\frac{1790}{1910}}\times\sqrt{\frac{1910}{1740}\times\frac{1790}{1800}}=\frac{1790}{1740}$$

It satisfies facto reversal test.

7.13.4 Circular Test

Circular test was given by Wester Guard. This test is like Time Reversal test but applied to more number of years. According to this test if data of the different periods is compared by shifting the base, we should be able to get the index of any period by correlating the different base periods used. Symbolically

$$=\frac{P_1}{P_0}\times\frac{P_2}{P_1}\times\frac{P_3}{P_2}=1$$

Only Simple Aggregative, Simple Geometric Mean of price relatives and Kelly's index meet this criterion.

TEST YOUR PROGRESS (B)

1. Find Laspeyre, Paasche and Fisher Index from following

Item	Price of the Base Year	Qty of the Base Year	Price of the Current Year	Qty of the Current Year
A	12	20	15	25
B	10	8	16	10
C	15	2	12	1
D	60	1	65	1
E	3	2	10	1

2. Calculate Laspeyre, Paasche, Bowley, Fisher, Marshal and Edgeworth price Index from following

Item	Qty of the Base Year	Expenditure of the Base Year	Qty of the Current Year	Expenditure of the Current Year
A	10	120	12	156
B	50	700	40	600
C	15	240	25	475
D	12	216	15	240

3. Calculate Laspeyre, Paasche, Fisher, Marshal and Edgeworth price Index from following

Item	Price 2015	Qty 2015	Price 2017	Qty 2017
A	5	100	6	150
B	4	80	5	100
C	2.5	60	5	72
D	12	30	9	33

4. Calculate index by using Weighted price relative method

Item	Price 2015	Price 2017	W

A	10	12	10
B	15	19	15
C	20	25	8
D	25	28	12

5. Apply Laspeyre, Paasche and Fisher Method to the following data and check whether these methods satisfy Time Reversal and Factor Reversal Test or not

Item	P_0	Q_0	P_1	Q_1
A	5	15	5	5
B	7	5	4	3
C	8	6	6	10
D	3	8	3	4

6. From the following data find out consumer price index number for the year 2020 taking 2018 as base by using (i) the aggregate expenditure method, and (ii) the family budget Method

Commodities	Quantity 2018	Price in 2018 (Rs.)	Price in 2020 (Rs.)
A	100	8	12
B	25	6	7.5
C	10	5	5.25
D	20	48	52
E	25	15	16.5
F	30	9	27

Answers:

1. Laspeyre - 129.09, Paasche – 130.13, Fisher 129.61
2. L = 106.35, P = 107.06, B = 106.72, F = 106.75, M & E = 106.7
3. L = 98.05, P = 99.18, F = 98.61, M & E = 98.68
4. 120.97,
5. L= 85.165, P = 86.232, F = 85.697, only Fisher method satisfies both tests.
6. 142.13

7.14 SUM UP

- Time series analysis is a situation where there are two variables in the problem and out of that one variable is necessarily the time factor.
- This analysis is very useful tool for forecasting.
- With passage of time there are fluctuations in the items.
- These fluctuations are mainly due to four factors called components of time series.
- These components are Secular trends, seasonal variations, cyclical variations and irregular variations.

- There are two models of time series, these are additive models and multiplicative models.
- For finding secular trends we can apply four methods the free-hand graphic method, the average Method, Moving Average Method and Method of Least Square.
- Index number shows change in a variable over a period of time.
- Price index shows change in price in current year in comparison to base year.
- Normally the base of index is taken as 100.
- There are different types of indexes like price index, quantity index, value index.
- Index number can be prepared without assigning weights or after assigning weights.
- Popular weighted aggregative index is Laspeyre, Paasche, Bowley, Fisher, Marshal Edgeworth and Kelly.
- Only Fisher index satisfies Time Reversal and Factor Reversal tests.

7.15 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define time series.

Q2. Types of time series.

Q3. What is multiplicative model?

Q4. Steps of moving average method.

Q5. Explain Index Number.

Q6. Give the names of Methods of index number.

Q7. Price Index

Q8. Quantity Index

Q9. Formula of fisher method?

Q10. Define Time Reversal.

Q11. Explain Factor Reversal Test.

Q12. What is Simple Price Relative Index numbers?

Q13. Explain Weighted Aggregative Index numbers.

Q14. Explain Simple Aggregative Index numbers.

Q15. What is Weighted Price Relative Index numbers?

B. Long Answer Type Questions

Q1. What is time series? Give its significance and limitations.

Q2. What are components of time series?

Q3. Give different types of trends in time series.

Q4. Give multiplicative and additive models of time series.

Q5. What is free hand curve method?

Q6. What is semi average method of time series?

Q7. How predictions are made using method of least square.

Q8. What is moving average trend. How it is determined.

Q9. Give various methods of estimating trends along with their respective merits and limitations.

Q10. What are index numbers? What are its uses?

Q11. Explain problems faced in construction of index numbers.

Q12. What are different types of Index numbers?

Q13. Explain different steps in construction of index numbers.

Q14. What are different methods of construction of index numbers?

Q15. What are tests of consistency of index numbers? Give various tests of consistency.

Q16. Explain the Time Reversal and Factor Reversal Test.

Q17. Why Fisher's Index is known as Ideal Index Number.

7.16 SUGGESTED READINGS

- J. K. Sharma, *Business Statistics,* Pearson Education.
- S.C. Gupta, *Fundamentals of Statistics,* Himalaya Publishing House.
- S.P. Gupta and Archana Gupta, *Elementary Statistics,* Sultan Chand and Sons, New Delhi.
- Richard Levin and David S. Rubin, *Statistics for Management,* Prentice Hall of India, New Delhi.
- M.R. Spiegel, *Theory and Problems of Statistics,* Schaum's Outlines Series, McGraw Hill, Publishing Co.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 8: PROBABILITY AND PROBABILITY RULES PROBABILITY DISTRIBUTIONS

STRUCTURE

8.0 OBJECTIVES

After reading this Unit, the learner should be able to know about:

- Elementary terms used in probability
- Conditional Probability
- Multiplication Law of Probability
- Independent Events
- Multiplication Law of Probability for Independent Events
- Discrete Probability Distributions
- Continuous Probability Distributions

8.1 INTRODUCTION

Probability plays one of a significant part in statistics. It measures the likelihood of an event. One can observe different practical situations where prediction is quoted. For instance, weather forecast quotes that there is 90% chance of rain today. Probability helps us to take decisions in uncertainty situations. It has important applications in all disciplines such as physics, chemistry, education, economics, etc. The probability theory has the purpose of providing mathematical models of situations affected or even governed by chance effects. One cannot predict with complete certainty the occurrence of the outcome of interest in any of the experiments. Broadly in theory of probability there are three possible states of expectation which are certainty, impossibility and uncertainty. The probability theory describes certainty by 1, impossibility by 0 and uncertainty lies between 0 and 1.

8.2 PREREQUISITES

1. **Set theory:** A set is a well-defined collection or aggregate of objects having given properties and specified according to a well-defined rule. The objects comprising the set are known as its elements.
   For e.g. A= set of first 10 natural numbers = {1,2,3,4,5,6,7,8,9,10} = $\{x, x \in N, x \leq 0\}$.
2. **Nulls set:** A set having no element at all is called a null or an empty set. It is represented by $\emptyset$. For e.g., if two dice are thrown and A is a set of points on the two dice so that their sum is greater than 12, then A is a null set.
3. **Subset:** A set A is said to be a proper subset of B if every element of A is also an element of Band there is at least one element of B which is not an element of A. We represent it by A$\subset B$.
4. **Equality of two sets:** Two sets A and B are said to be equal, if every element of A is an element of B and if every element of B is an element of A. We represent it as
$$A = B \ \ if\ x \in A \ \rightarrow x \in B \ \ and\ x \in B \rightarrow x \in A.$$

Remark-

- Every set is a subset of itself *i.e.,* $A \subset A$.
- The null set $\emptyset$ is a subset of every set

5. **Universal set**: The overall limiting set of which all the sets under consideration are subsets
6. **Union of two sets:** If A and B are two sets then their union is defined as a set of elements that belong to either A or B or both. It is represented as
$$A \cup B = \{x \in A\ or\ x \in B\}.$$
7. **The intersection of two sets:** defined as a set whose elements belong to both A and B. Symbolically it is written as
$$A \cap B = \{x \in A\ and\ x \in B\}.$$
8. **Difference of two sets**: The difference between sets A and B is defined as

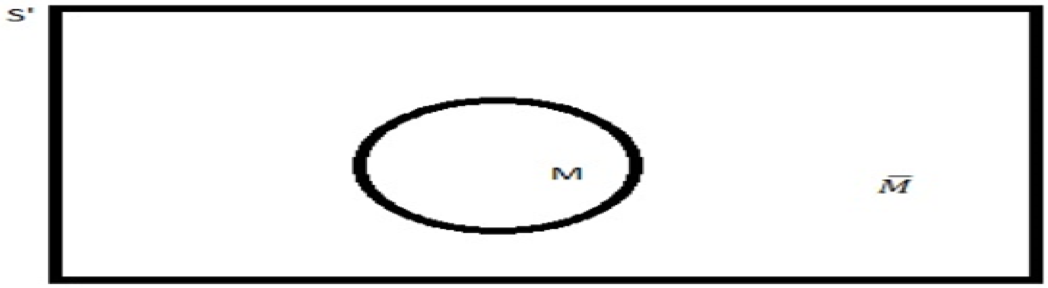
$A - B = \{x | x \in A\ and\ x\ doesnot\ belong\ to\ B\}$.

For e.g. A= {1,3,5,7,9,11,13}, B= {5,9,13,15}, A-B= {1,3,711}.

Basics of Probability

1. **Random experiment**: An experiment that is performed under necessary homogeneous conditions.

2. **Trial**: Trial is defined as when a random experiment is carried.
3. **Event/Outcome**: An outcome of a random experiment is known as an event. For example, a coin tossed in the air is an experiment and a coin land with a head up is the outcome of an experiment. A parachute jumps from a plane is an experiment and it will fall down in its outcome. The outcome of a random experiment which entails the occurrence of an event A is known as favorable outcome to A.
4. **Exhaustive cases**: The total number of possible outcomes of a random experiment is called exhaustive cases. For e.g. In tossing a dice total no. of outcomes is 6 i.e., 1,2,3,4,5,6.
5. **Favourable cases**: The total number of outcomes of a random experiment that confirms the happening of an event.
6. **Mutually Exclusive cases**: Any two or more events are called mutually exclusive if the occurrence of one of the events doesn't affect the occurrence of the other. For e.g. In throwing of a dice, the set of all possible outcomes is mutually exclusive.
7. **Equally probable cases**: The possible outcomes are said to be equally likely if none of the cases is preferred as compared to others. For instance, in tossing of a coin the occurrence of head and tail are equally probable.
8. **Independent events:** Cases are said to be independent if the occurrence of one event is not affected by the occurrence of the other event. For e.g., picking two balls from a bag containing 'a 'red balls and 'b' white balls is a trial and occurrence of both red balls, both white balls and one red and second ball white are independent events.
9. **Simple event:** An event that includes one and only one of the final outcomes for a random experiment is called a simple event.
10. **Compound event:** A compound event consists of more than one outcome.
11. **Complementary event:** Let 'M' represent the event of the number of favorable responses in the experiment and $\overline{M}$ represent the complementary event. It is defined as the number of non-favorable cases in the experiment. The representation of an event and its complement in the Venn diagram is discussed below:



For example, in a group of 2000 taxpayers, 400 have been audited by the IRS at least once.

If one taxpayer is randomly selected from this group. The two complementary events for this experiment and their probability are

P(M) = 400/2000

P $(\overline{M}) = 1600/2000$

Where 'M' represents a taxpayer, who has been audited by the IRS at least once

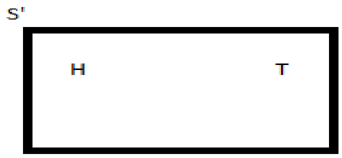
$\overline{M}$ is the selected taxpayer never been audited

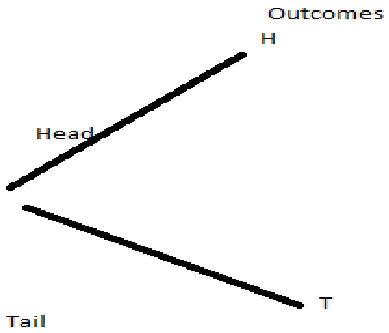
i) Odd in favor: Let $P = A/N$ is the probability that the event 'M' occurs and $Q = B/N$

be the probability that the event 'M' does not occur where $A + B = N$. If $p \geq 1/2$ then odd in favor of 'M' is the ratio A:B

ii) Odd against favor: If $p \leq 1/2$, odd against A is defined as B: A

12. **Venn diagram:** It is a pictorial representation in the form of rectangle, square or circle that predicts all the possible outcomes of an experiment
13. **Tree diagram:** It is the diagram in which each outcome is represented by a branch of a tree. For example, The Venn and tree diagram of tossing a coin once if given as





where sample space = {H, T}, H represents head, T as tail.

14. **Intersection of events**: It gives us the outcomes that are common
15. **Joint Probability**: The probability of the intersection of two events is called joint probability P (M and N)
16. **Sampling with replacement**: In this object was drawn at random is placed back into the given set and the set is mixed thoroughly. Then we draw the next object at a random.
17. **Sampling without replacement**: In this the object that was drawn is put aside.

8.3 CLASSICAL DEFINITION OF PROBABILITY

If there are n mutually exclusive, equiprobable and exhaustive elements in a sample space S, and if r of them are favourable to the occurrence of some event A, then the probability of event A is given by the formula:

$$P(A) = \frac{\text{Number of Favourable Cases to } A}{\text{Total Number of Cases}}$$

$$= \frac{r}{n}$$

The favourable cases to happening of A are always less than or equal to the totalexhaustive cases, and also, they cannot be negative.

Therefore, $0 \leq r \leq n$. Dividing by n, we have

$$0 \leq \frac{r}{n} \leq 1, \text{ i.e., } 0 \leq P(A) \leq 1.$$

If r cases are favourable to the happening of an event A, then $n - r$ cases are favourable to not happening of A. Therefore,

$$P(A) = \frac{\text{Number of Favourable Cases to } A}{\text{Total Number of Cases}}$$

$$= \frac{n-r}{n}$$

$$= 1- \frac{r}{n}$$

$$= 1 - P(A)$$

Thus, we have $P(A^c) = 1 - P(A)$.

8.4 STATISTICAL OR EMPIRICAL DEFINITION OF PROBABILITY

If the random experiment is repeated under essentially the same conditions for a large number of times, then the limit of the ratio of number of times an event happens to the total number of trials is defined as the probability of that event. Here assume that the limit exists.

$$P\,(A) = \frac{lim}{n\to\infty}\,\frac{r}{n}$$

8.5 THEOREMS ON PROBABILITY

8.5.1 Probability of Impossible Event

The probability of an impossible event is always zero, i.e., P(ϕ) =0.

Proof: Let ϕ be an impossible event of sample space S. Then, we have $\phi \cup S = S$ and $\phi \cap S = \phi$. Thus, ϕ and S are mutually exclusive and exhaustive events. Hence by Axiom 2 and3,

$P\ (\emptyset \cup S) = P(S)$

$P(\emptyset) + P(S) = P(S)$, (Using Axiom 3)

$P(\emptyset) + 1 = 1$(Using Axiom 2)

i.e., $P(\emptyset) = 0$.

8.5.2 Probability of Complementary Event

If A and A^c are complementary events of the sample space S, then $P(A)+ P(A^c) = 1$.

Proof

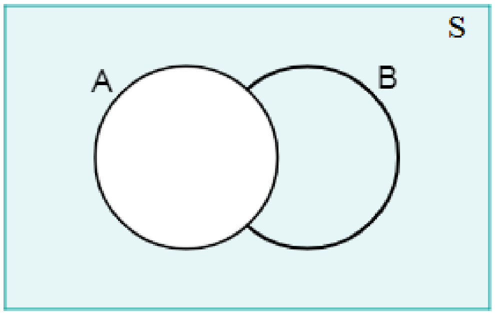


Figure: Complementary Event of *A*

Since A and A^c are complementary events of the sample space S, we have $A \cup A^c = S$ and

$A \cap A^c = \varphi$. Hence by Axiom 2 and 3,

$$P(A \cup A^c) = P(S)$$

$$P(A) + P(A^c) = P(S), \text{(Using Axiom 3)}$$

$$P(A) + P(A^c) = 1\text{(Using Axiom 2)i.e., } P(A^c) = 1 - P(A).$$

Thus, we have $P(A)+ P(A^c) = 1$.

8.5.3 Addition Law of Probability

If A and B are two events of non-empty sample space S, then

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof

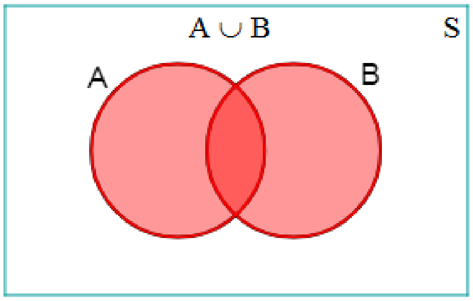


Figure 7.2: Union of A and B

The above Venn diagram shows $A \cup$ B.

From the Venn diagram in Figure 7.3, it is clear that the event $A \cup$ B can be written as $A \cap$ B = $A \cup (A^c \cap B)$ such that A and $A^c \cap B$ are mutually exclusive.

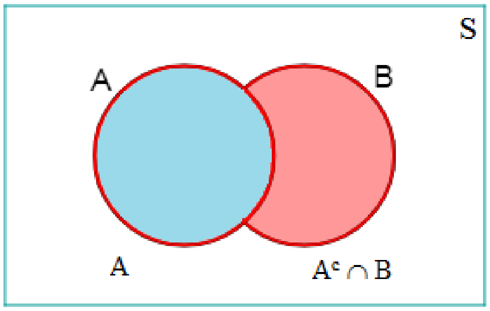


Figure 7.3: $A \cup B = A \cup (A^c \cap B)$

Using Axiom 3,

$$P(A \cup B) = P[A \cup (A^c \cap B)]$$

$$= P(A) + P(A^c \cap B) \qquad \ldots\ldots (1)$$

Similarly, from Figure 7.4, the event B can be written as a union of two mutuallyexclusive events $A \cap B$ and $A^c \cap B$ such that $B = (A \cap B) \cup (A^c \cap B)$.

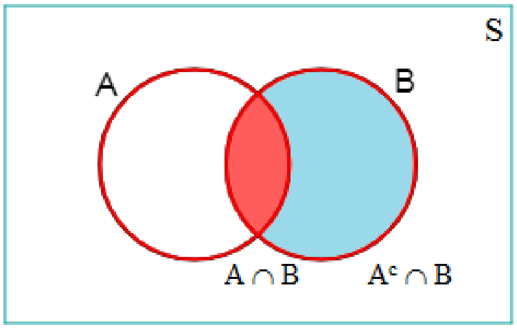


Figure 7.4: $B = [(A \cap B) \cup (A^c \cap B)]$

$$P(B) = P[(A \cap B) \cup (A^c \cap B)]$$

$$= P(A \cap B) + P(A^c \cap B)$$

$$\therefore P(A^c \cap B) = P(B) - P(A \cap B)$$

Substituting for $P(A^c \cap B)$ in equation 1, we get

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

8.5.4 Addition Law of Probability for Mutually Exclusive Events

If A and B are two mutually exclusive events of non-empty sample space S, then

$$P(A \cup B) = P(A) + P(B).$$

Proof: Since A and B are mutually exclusive events, $A \cap B = \phi$.

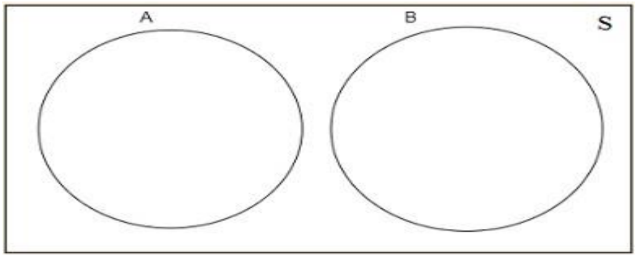


Figure 7.5: Mutually Exclusive Events

Therefore, $P\ (A\cap B) = P(\phi) = 0$. Hence,

$P(A \cup B) = P(A) + P(B)$

8.5.5 Law of Addition for Three Events

For three events A, B and C of non-empty sample space S,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C)$$
$$-P(B \cap C) + P(A \cap B \cap C)$$

<u>Example 1</u>: Consider a random experiment of throwing a fair die. List all the elements of thesample space. Find the probability that

(i) Even number appear on the top

(ii) Odd number appears on the top

(iii) A number greater than or equal to 3 appear on the top.

Solution**:** The sample space S of the random experiment is S = {1, 2, 3, 4, 5, 6}.

(i) Let A be the event that even number appear on the top. Then A = {2, 4, 6}.

$$\mathrm{P}(A) = \frac{\text{Number of favourable cases}}{\text{Total number of cases}} = \frac{3}{6} = \frac{1}{2}$$

(ii) Let B be the event that odd number appear on the top. Then B = {1, 3, 5}

(iii) $\mathrm{P(B)} = \frac{\text{Number of favourable cases}}{\text{Total number of cases}} = \frac{3}{6} = \frac{1}{2}$

(iv) Let C be the event that a number greater than or equal to 3 appears on the top

Then C = {3, 4, 5, 6}

$$\mathrm{P(C)} = \frac{\text{Number of favourable cases}}{\text{Total number of cases}} = \frac{4}{6} = \frac{2}{3}$$

<u>Example 2</u>: Consider a random experiment of tossing three coins. List all the elements of thesample space. Find the probability of getting

(i) at least one head,

(ii) at the most two heads,

(iii) no head.

Solution**:** Let three coins be tossed. The sample space S consists of 8 elementary eventsS = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

(i) Let A denote the event that at least one head occurs.

Then A = {HHH, HHT,HTH, THH, HTT, THT, TTH}.

$$\mathrm{P}(A) = \frac{\text{Number of favourable cases to A}}{\text{Total number of cases}} = \frac{7}{8}$$

(ii) Let B denote the event that at the most two head occur.

Then B = {HHT, HTH,THH, HTT, THT, TTH, TTT}

$$\mathrm{P(B)} = \frac{\text{Number of favourable cases to B}}{\text{Total number of cases}} = \frac{7}{8}$$

(iii) Let C denote the event that no head occur.

Then C ={TTT}

$$\mathrm{P(C)} = \frac{\text{Number of favourable cases to C}}{\text{Total number of cases}} = \frac{1}{8}$$

Example 3: In a single throw with two uniform dice, what is the probability of getting

(a) a total of 9,

(b) total different from 9,

(c) total is greater than or equal to 8,

(d) a total of 7 or 11,

(e) maximum of two numbers is greater than 4.

Solution: Let two uniform dice be thrown. The sample space consists of 36 elementary events.

S = {(1,1), (1,2), (1,3), (1,4), (1,5), (1,6),

(2,1), (2,2), (2,3), (2,4), (2,5), (2,6),

(3,1), (3,2), (3,3), (3,4), (3,5), (3,6),

(4,1), (4,2), (4,3), (4,4), (4,5), (4,6),

(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),

(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}.

(a) Let A be the event that a total of 9 occur. Therefore A = {(3, 6), (4, 5), (5, 4), (6,3)}.So, no. of elements in A is r = 4. Hence, the required probability is

$$\text{P(B)} = \frac{\text{Number of favourable cases A}}{\text{Total number of cases}} = \frac{4}{36} \;\; = = \frac{1}{9}$$

(b) The event of getting totally different from 9 is the complementary event of A. So,the required probability is

$$P(A^c) = 1 - P(A) = 1 - \frac{1}{9}$$

(C) Let B be the event that a total of greater than or equal to 8 occurs. Therefore,

B = { (2,6),

(3,5), (3,6),

(4,4), (4,5), (4,6),

(5,3), (5,4), (5,5), (5,6),

(6,2), (6,3), (6,4), (6,5), (6,6)}.

Number of favorable cases for B = 15. Therefore

P (B) = 15

$$\frac{\text{Number of favourable cases to B}}{\text{Total number of cases}} = \frac{15}{36}$$

—

(d) Let C be the event that total of 7 occur and D be the event that the total of 11 occur.

C = {(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)} and

D = {(5,6), (6,5)}.

The required probability is C ∪ D.

Moreover, C and D are mutually exclusive.

$$P(C \cup D) = P(C) + P(D) = \frac{6}{36} + \frac{2}{36} \;\; = \frac{8}{36} \;\; = \frac{2}{36}$$

(e) Let E = {(x,y) | max (x,y) >4}. Then

E = { (1,5), (1,6)

(2,5), (2,6),

(3,5), (3,6),

(4,5), (4,6),

(5,1), (5,2), (5,3), (5,4), (5,5), (5,6),

(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)}.

No. of cases favourable to E = 20. Therefore, the probability of E is

$P(E) = \frac{20}{36} = \frac{5}{36}$ — –

<u>Example 4</u>: In a lot of 100 electric bulbs 10% of them are defective. Five bulbs are selected atrandom.

a. **What is the probability of no defective bulb among the 5 bulbs?**
b. **What is the probability of 2 bulbs being defective among the 5 bulbs?**

Solution**:** Total number of bulbs = 100. Out of 100 electric bulbs, 10% of bulbs are defective, i.e., 10 bulbs are defective and remaining 90 are non-defective.

Out of 100 bulbs 5 bulbs can be selected in $(\frac{100}{5})$ ways.

i.e., n = Total number of elements = $\frac{100}{5}$

And out of 90 non-defective bulbs 5 can be selected in $\frac{90}{5}$ ways. i.e., m = $\frac{90}{5}$

a. Let A be the event that no defective bulb among the selected 5.

$$P(A) = \frac{m}{n} \qquad = \frac{\frac{90}{5}}{\frac{100}{5}}$$

b. Let B be the event that two defective bulbs among selected 5.

$$P(A) = \frac{m}{n} \qquad = \frac{\frac{90}{5}}{\frac{100}{5}}$$

Out of 10 defective bulbs 2 bulbs can be selected in $(\frac{10}{2})$ ways and remaining 3 bulbs from 90 non-defective can be selected in $(\frac{90}{3})$ ways. The events are compound events.

P (B) = $\frac{90}{3} = \frac{(\frac{10}{2})(\frac{90}{3})}{\frac{100}{5}}$

<u>Example 5</u>: Two cards are drawn at random simultaneously from a pack of playing cards. Find theprobability that

- **a. both the cards are spade cards,**
- **b. both the cards are of same suit.**

Solution**:** As two cards are drawn from 52 cards, the sample space consists of $(\frac{\mathbf{52}}{\mathbf{2}})$ elements.

a. Let A be the event that both the cards are space cards. Out of 13 cards of same suit,2 cards can be selected in $(\frac{13}{2})$ ways.

P (A) = $\frac{\frac{13}{2}}{\frac{52}{2}} = \frac{13*12}{52*51} = \frac{1}{17}$Let B be the event that both the cards are of same suit. Two cards of any suit can be selected in $(\frac{13}{2})$ ways, but there are 4 suits. So, the number of elements favourable to B = 4 × $(\frac{13}{2})$.

P (B) = $\frac{4*\frac{13}{2}}{\frac{52}{2}} = \frac{4*13*12}{52*51} = \frac{4}{17}$

<u>8.6 Conditional Probability</u>

In a simple language the conditional probability is "What is the chance that somethingwill happen, given that something else has already happened?".

Let *A* and *B* be two events of non-empty sample space *S*.

The probability of some event *A* when it is known that event *B* has already occurred iscalled conditional probability of *A* given *B*. It is denoted by *P(A|B)* and is defined by

P (A|B) = $\frac{P\ (A \cap B)}{P\ (B)}$ = P(B) > 0

P(A|B) can be read as *probability that A occurs given that B occurs*. The conditional probability of *A* given *B* is the joint probability of *A* and *B* divided by the marginal probability of *B*.

The probability of some event *B* when it is known that event *A* has already occurred is called conditional probability of *B* given *A*. It is denoted by *P(B|A)* and is defined by

P (B|A) = $\frac{P\ (A \cap B)}{P\ (B)}$ = P(A) > 0

P(B|A) can be read as *probability that B occurs given that A occurs*. The conditional probability of *B* given *A* is the joint probability of *B* and *A* divided by the marginal probability of *A*.

8.7 Multiplication Law of Probability

Let *A* and *B* be two events of non-empty sample space *S* such that *P(A) > 0*, *P(B) > 0*,then and That is,

$P(A \cap B) = P(A|B) \cdot P(B)$

and

$$P(B \cap A) = P(B|A) \cdot P(A).$$

That is,

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Multiplication law of probability is also known as *Law of Compound Probability*.

Example 6: A random sample of 200 students is classified below by sex and experiencing hypertension during the examination period. If a student is selected at random from this sample, find the probability that the student is

(a) male given that the student is experiencing hypertension,

Status \Gender	Male	Female	Total
Hypertension	60	45	105
No Hypertension	40	55	95
Total	100	100	200

(b) experiencing hypertension given that the student is female,

(c) female given that the student is experiencing no hypertension.

Solution: Let *M* be the event that male student is chosen, *F* be the event that female students is chosen, *H* be the event that the one chosen experiences hypertension and *N* be the event that the one experiences no hypertension.

(a) Probability that the student is male given that he is experiencing hypertension is

$$P(M|H) = \frac{P\ (M \cap\ H)}{P(H)}$$

$$= \frac{60/200}{105/200} = \frac{60}{105}$$

(b) Probability that the student is experiencing hypertension given that the student isfemale

$P(\text{H}|\text{F}) = \frac{P\ (H \cap F)}{P(F)}$

$= \frac{45/200}{100/200} = \frac{45}{100}$

(c) Probability that the student is female given that she is experiencing no hypertension is

$P(M|H) = \frac{P\ (F \cap\ N)}{P(N)}$

$= \frac{55/200}{95/200} = \frac{55}{95}$

8.8 INDEPENDENT EVENTS

Events are said to be independent if occurrence or non-occurrence of any one of themdoes not depend on that of any of the remaining ones.

In particular, an event A is said to be independent of another event B if $P(A|B) = P(A)$.This definition is meaningful only if $P(A|B)$ is defined, i.e., $P(B) > 0$.

8.9 Multiplication Law of Probability For Independent Events

If A and B are events such that $P(A) > 0$, $P(B) > 0$, then A and B are independent if andonly if

$P\ (A \cap B) = P(A) \cdot P(B)$.

i.e., for independent events A and B, the probability that both of these occursimultaneously is the product of their respective probabilities.

Note: Three events A, B and C of the sample space S, are mutually independent if

a. $P\ (A \cap B) = P(A)P(B)$

b. $P\ (B \cap C) = P(B)P(C)$

c. $P\ (A \cap C) = P(A)P(C)$

d. $P\ (A \cap B \cap C) = P(A)P(B)P(C)$

If only first three conditions are satisfied, then A, B and C are pair-wise independent.

Example 7: Jaydev and Vijay can solve respectively 60% and 80% problems in a book. They try independently to solve a problem randomly selected from the book. Find the probability that (i) problem is solved, (ii) only Jaydev can solve the problem, (iii) only Vijay can solve the problem, (iv) none of them can solve the problem.

Solution: Let A be the event that Jaydev can solve the problem, and B be the event that Vijay can solve the problem. Hence $P(A) = 0.60$ and $P(B) = 0.80$.

As Jaydev and Vijay solve the problem independently, we have

$$P\ (A \cap B) = P(A) \cap P(B) = 0.60 \cap 0.80 = 0.48.$$

(i) The problem is solved if both Jaydev and Vijay solve the problem. Thus, therequired probability is

$$\begin{aligned} P\ (\text{Problem is solved}) &= P\ (A \cup B) \\ &= 1 - P\ (A^c \cap B^c) \\ &= 1 - P(A^c)P(B^c) \\ &= 1 - (0.40)\ (0.20) \\ &= 1 - 0.08 = 0.92 \end{aligned}$$

(ii) The probability that only Jaydev can solve the problem is $P\ (A \cap B^c)$.

$P(A \cap B^c) = P(A \cap B^c)$

$= P(A)P(B^c) \quad = (0.60)\ (0.20)= 0.12$

(iii) The probability that only Vijay can solve the problem is $P(A^c \cap B)$.

$P(A^c \cap B) = P(A^c \cap B)$

$= P(A^c)P(B) \quad = (0.40)\ (0.80)= 0.32$

(iv) The probability that none can solve the problem is $P(A^c \cap B^c)$.

$P(A^c \cap B^c) = P(A^c \cap B^c)$

$= P(A^c)P(B^c) \quad = (0.40)\ (0.20)= 0.08$

8.10 MEANING OF PROBABILITY DISTRIBUTIONS

A probability distribution is a mathematical function that describes the likelihood of various outcomes or events in a random experiment or process. It provides a way to represent and quantify uncertainty or randomness. In other words, it tells you how the possible values of a random variable are distributed or spread out.

There are two main types of probability distributions:

- Discrete Probability Distribution: This type of distribution deals with random variables that can only take on distinct, separate values. Examples of discrete probability distributions include the Bernoulli distribution (which models a single trial with two possible outcomes, like success or failure) and the Poisson distribution (which models the number of events occurring in a fixed interval of time or space).
- Continuous Probability Distribution: This type of distribution deals with random variables that can take on any value within a continuous range. Examples of continuous probability distributions include the normal distribution (often referred to as the Gaussian distribution), the exponential distribution, and the uniform distribution.

Probability distributions are fundamental in statistics and probability theory. They are used in various fields, such as science, engineering, finance, and data analysis, to model and understand random processes and make predictions or inferences based on uncertainty. Different distributions are used to model different types of data, and the choice of distribution depends on the characteristics of the data and the problem being analysed.

8.11 DIFFERENT TYPES OF PROBABILITY DISTRIBUTIONS

We have seen what Probability Distributions are, now we will see different types of Probability Distributions. The Probability Distribution's type is determined by the type of random variable. There are two types of Probability Distributions:

- Discrete Probability Distributions for discrete variables
- Cumulative Probability Distribution for continuous variables

Discrete Probability Functions also called Binomial Distribution assume a discrete number of values. For example, coin tosses and counts of events are discrete functions. These are discrete distributions because there are no in-between values. We can either have heads or tails in a coin toss. For discrete probability distribution functions, each possible value has a non-zero probability. Moreover, the sum of all the values of probabilities must be one.

For example, the probability of rolling a specific number on a die is 1/6. The total probability for all six values equals one. When we roll a die, we only get either one of these values.

8.11.1 Binomial Distribution

It is a random variable that represents the number of successes in “N” successive independent trials of Bernoulli’s experiment. It is used in a plethora of instances including the number of heads in “N” coin flips, and so on. Let P and Q denote the success and failure of the Bernoulli Trial respectively. Let’s assume we are interested in finding different ways in which we have 1 success in all six trials. There are six cases are available as listed below:

PQQQQQ, QPQQQQ, QQPQQQ, QQQPQQ, QQQQPQ, QQQQQP

Likewise, 2 successes and 4 failures will show combinations thus making it difficult to list so many combinations. Henceforth, calculating probabilities of 0, 1, 2…, n number of successes can be long and time-consuming. To avoid such lengthy calculations along with a listing of all possible cases, for probabilities of the number of successes in n-Bernoulli’s trials, a formula is made which is given as:

If Y is a Binomial Random Variable, we denote this Y∼ Bin (n, p), where p is the probability of success in a given trial, q is the probability of failure, let ‘n’ be the total number of trials, and ‘x’ be the number of successes, the Probability Function P(Y) for Binomial Distribution is given as:

$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$

Or

$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (q)^{n-x}$

n = the number of experiments

x = 0, 1, 2, 3, 4, …

p = Probability of Success in a single experiment

q = Probability of Failure in a single experiment = 1 – p

Properties of Binomial Distribution

- There are two possible outcomes: true or false, success or failure, yes or no.
- There is ‘n’ number of independent trials or a fixed number of n times repeated trials.
- The probability of success or failure remains the same for each trial.
- Only the number of successes is calculated out of n independent trials.
- Every trial is an independent trial, which means the outcome of one trial does not affect the

outcome of another trial.

Example 8: If a coin is tossed 5 times, find the probability of:

(a) Exactly 2 heads

(b) At least 4 heads.

Solution: (a) The repeated tossing of the coin is an example of a Bernoulli trial. According to the problem: Number of trials: n=5

Probability of head: p= 1/2 and hence the probability of tail, q =1/2

For exactly two heads: x=2

$P(x=2) = {}^5C_2\ p^2\ q^{5-2} = 5!\ /\ 2!\ 3! \times (\frac{1}{2})^2 \times (\frac{1}{2})^3$

$= \frac{5*4}{1*2} \times (\frac{1}{2})^2 \times (\frac{1}{2})^3 \qquad = 10*(\frac{1}{4})*(\frac{1}{8})$

$P(x=2) = \frac{5}{16}$

(b) For at least four heads,

$x \geq 4$, $P(x \geq 4) = P(x = 4) + P(x=5)$

Hence, $P(x = 4) = {}^5C_4\ p^4\ q^{5-4} = \frac{5}{1} \times (\frac{1}{2})^4 \times (\frac{1}{2})^1 = \frac{5}{32}$

$P(x = 5) = {}^5C_5\ p^5\ q^{5-5} = (\frac{1}{2})5 = \frac{1}{32}$

Therefore, $P(x \geq 4) = 5/32 + 1/32 = 6/32 = 3/16$

Example 9: In a game of darts suppose you have a 25% chance that you will hit the bullseye. If you take a total of 15 shots then what is the probability that you will hit the bullseye 5 times?

Answer: n = 15, p = 25 / 100 = 0.25, x = 5

We have to use the Binomial probability distribution given by

$P\ (X = x) = {}^nC_x\ p^x(1-p)^{n-x}$

$P\ (X = 5) = ({}^{15}C_5)0.25^5(1-0.25)^{15-5} = 0.165$

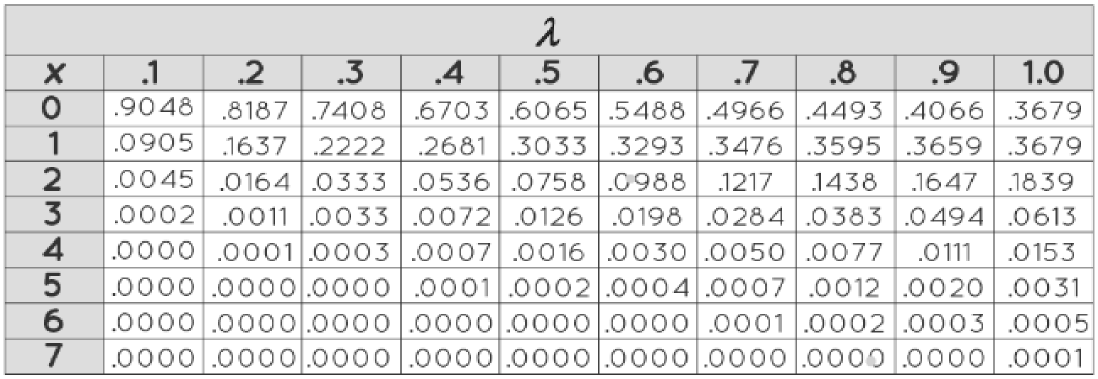
8.11.2 Poisson Distribution

The Probability Distribution of the frequency of occurrence of an event over a specific period is called Poisson Distribution. It tells how many times the event occurred over a specific period. It basically counts the number of successes and takes a value of the whole number i.e. (0,1,2…). It is expressed as

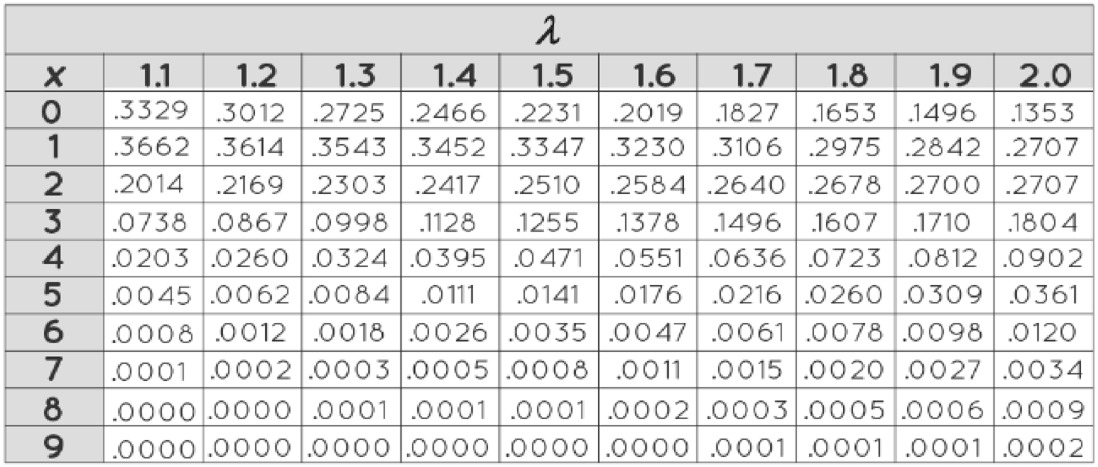
$f(x) = P(X=x) = (e^{-\lambda}\ \lambda^{x}\ )/x!$

where, **x** is number of times event occurred, λ: mean number of successes that occur during a specific interval, k: number of successes, e: a constant equal to approximately 2.71828.

Poisson Distribution Table



	λ									
x	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001



	λ									
x	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

A Poisson experiment is an experiment that has the following properties:

- The number of successes in the experiment can be countable.
- The mean number of successes that occurs during a specific interval of time (or space) is known.
- Each outcome is independent.

- The probability that a success will occur is proportional to the size of the interval.

Example 10: Find the mass probability of function at x = 6, if the value of the mean is 3.4.

Answer: Given: $\lambda = 3.4$, and $x = 6$.

Using the Poisson distribution formula:

$P(X = x) = (e^{-\lambda} \lambda^{x})/x!$

$P(X = 6) = (e^{-3.4} 3.4^{6})/6!$

$P(X = 6) = 0.072$

The probability of function is 7.2%.

8.12 CUMULATIVE PROBABILITY DISTRIBUTION (NORMAL DISTRIBUTION)

Cumulative Probability Distribution takes value in a continuous range. An example, the range may consist of a set of real numbers. In this case, Cumulative Probability Distribution will take any value from the continuum of real numbers unlike the discrete or some finite value taken in the case of Discrete Probability distribution**.** Cumulative Probability Distribution is of two types, Continuous Uniform Distribution, and Normal Distribution. Continuous probability functions are also known as probability density functions. You know that you have a continuous distribution if the variable can assume an infinite number of values between any two values. Continuous variables are often measurements on a scale, such as height, weight, and temperature.

Unlike discrete probability distributions where each particular value has a non-zero likelihood, specific values in continuous probability distribution functions have a zero probability. For example, the likelihood of measuring a temperature that is exactly 32 degrees is zero. Continuous Uniform Distribution is described by a density function that is flat and assumes value in a closed interval let's say [P, Q] such that the probability is uniform in this closed interval. It is represented as f (x; P, Q).

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values in a continuous range. Some more examples are like, number of customers arriving at a salon in an hour, number of suicides reported in a particular city, number of printing errors on each page of the book.

Properties of Normal Distribution

- For the normal distribution of data, the mean, median, and mode is equal.

i.e., Mean = Median = Mode

- Total area under the normal distribution curve is equal to 1.
- It is a Unimodal Curve, i.e., a curve with only one peak.
- Normal Distribution Curve is always bell-shaped.
- It is symmetric at the center along the mean.
- In a normally distributed curve, there is exactly half value to the right of the central and exactly half value to the right side of the central value.
- It is defined with the values of the mean and standard deviation.

The formula for the normal distribution is;

$$\mathrm{P(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where, x is the variable, μ is the mean, σ is the standard deviation.

Example 11: Calculate the probability density function of normal distribution using the following data. x = 3, μ = 4 and σ = 2.

Solution: Given, variable, x = 3

Mean = 4 and

Standard deviation = 2

By the formula of the probability density of normal distribution, we can write;

$$f(3, 4, 2) = \frac{1}{2\sqrt{2\pi}} e^{\frac{-(3-2)^2}{2\times 2^2}}$$

Hence, f (3,4,2) = 1.106.

8.13 SUM UP

In this unit we have introduced the concept of probability and various definitions of probability, how to compute the probabilities of events using different laws of probability.We have discussed the concept of conditional probability, multiplication law of probability and independence. To understand the theoretical concepts.

8.14 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define the law of Probability

Q2. Given the classical definition of probability. State its limitation.

Q3. State the addition law of probability for two events.

Q4. State the addition law of probability for three events.

Q5. What is the probability of impossible event?

Q6. What is the probability of certain event?

Q7. Determine the type of events in each case

a. $P\ (A \cap B) = 0$

b. $P\ (A \cap B) = P(A) + P(B) =1$

c. $P\ (A \cap B) = P(A) + P(B).$

Q8. Define conditional probability.

Q9. State the multiplication law of probability.

Q10. State the condition for independence of two events.

Q11. Explain the binomial distribution properties.

Q12. Explain the Poisson distribution with example.

Q13. Explain the Normal distribution with example.

B. Long Answer Type Questions

Q1. A die is thrown once. What is the probability that

a. a number 5 appear on the top,

b. an even number appear on the top,

c. an odd number appear on the top,

d. a number less than 0 appear on the top,

e. a number greater than or equal to 0 appear on the top.

Q2. Explain multiplication law of probability.

Q3. Two cards are drawn without replacement from a pack of 52 cards. What is the probability that (i) Both are drawn are red, (ii) First is king and second is queen, (iii) One is red and other is black?

Q4. A charted accountant applies for a job in two firms X and Y. He estimates that the probability of his being selected in firm X is 0.7, and being rejected at Y is 0.5, and the probability of at least one of his applications being rejected is 0.6. What is the probability that he will be selected in one of the firms?

Q5. Give the meaning and Properties of Discrete series**.**

Q6. What do you mean by normal distribution, give properties.

8.15 SUGGESTED READINGS

- Gupta, S.C. and Kapoor, V.K. (2014): Fundamentals of Mathematical Statistics,Sultan Chand & Sons, New Delhi, 12th Edition.
- Hastie, Trevor, et al. (2009): The Elements of Statistical Learning, Springer
- Ross, S.M. (2004): Introduction to Probability and Statistics for Engineers andScientists, Academic Press
- Navidi, W. (2011): Statistics for Engineers and Scientists, McGraw Hill, ThirdEdition.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

Unit 9: Tests of Hypothesis–I, Tests of Hypothesis – II, Chi-Square Test

STRUCTURE

9.0 OBJECTIVES

After reading this unit, learners will be able to learn about:

- Meaning of Hypothesis (Null, Alternative hypothesis)
- Basic terms used in hypothesis like acceptance and rejection region, power of test, test of significance, types of errors (Type I and Type- II)
- Parametric and Non-parametric tests
- Chi-Square test and its applications

9.1 INTRODUCTION/ MEANING OF HYPOTHESIS

A hypothesis is an assumption of the association between two or more variables. The population, the variables, and the relationships between the variables are necessary for the hypothesis to be complete. The hypothesis does not have to be correct. While the hypothesis forecasts what the researchers expect to see, the goal of the research is to determine whether this guess is right or wrong. When experimenting, researchers might explore several factors to determine which ones might contribute to the outcome. In many cases, researchers may find that the results of an experiment do not support the original hypothesis. When writing up these results, the researchers might suggest other options that should be explored in future studies.

9.2 BASIC CONCEPTS OF HYPOTHESIS

The hypothesis is a fundamental concept in the scientific method and research process. It serves as a starting point for investigations and experiments, guiding researchers in their pursuit of knowledge. Here are some basic concepts related to hypotheses.

9.2.1 Null Hypothesis

A null hypothesis is a hypothesis in which the sample observations result from a chance. It is said to be a statement in which the researcher wants to examine the data. It is denoted by H_0. In statistics, the null hypothesis is usually denoted by the letter H with subscript '0' (zero), such that H_0 (pronounced as H-null or H-zero or H-nought). At the same time, the alternative hypothesis expresses the observations determined by the non-random cause. It is represented by H_1 or Ha. The main purpose of a null hypothesis is to verify/ disprove the proposed statistical assumptions.

An example of the hypothesis is as, If the hypothesis is that, "If random test scores are collected from men and women, does the score of one group differ from the other?" a possible null hypothesis will be that the mean test score of men is the same as that of the women.

H_0: $\mu_1 = \mu_2$

H_0= null hypothesis

μ_1= mean score of men

μ_2= mean score of women

Sometimes the null hypothesis is rejected too. If this hypothesis is rejected means, that research (assumption) could be invalid. Many researchers will neglect this hypothesis as it is merely opposite to the alternate hypothesis. It is a better practice to create a hypothesis and test it. The goal of researchers is not to reject the hypothesis. However, a perfect statistical model is always associated with the failure to reject the null hypothesis.

9.2.2 Alternative Hypothesis

An alternative hypothesis is a statement that describes that there is a relationship between two selected variables in a study. An alternative hypothesis is usually used to state that a new theory is preferable to the old one (null hypothesis). This hypothesis can be simply termed as an alternative to the null hypothesis.

The alternative hypothesis is the hypothesis that is to be proved that indicates that the results of a study are significant and that the sample observation does not result just from chance but from some non-random cause.

If a study is to compare method A with method B about their relationship and we assume that method A is superior or method B is inferior, then such a statement is termed an alternative hypothesis. The symbol of the alternative hypothesis is either H1 or Ha while using less than, greater than, or not equal signs.

The following are some examples of alternative hypothesis:

If a researcher is assuming that the bearing capacity of a bridge is more than 10 tons, then the hypothesis under this study will be:

Null hypothesis H_0: $\mu = 10$ tons

Alternative hypothesis H_1: $\mu > 10$ tons

9.2.2.1 One-tailed & Two-tailed

A test of testing the null hypothesis is said to be a 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 a one-tailed test. For example, if our null and alternative hypothesis is $H_{0:}$ $\mu = \mu_0$ and H_1: $\mu \neq \mu_0$ then the test for testing the null hypothesis is two-tailed because the alternative hypothesis is two-tailed which means, the parameter μ can take value greater than μ0 or less than μ_0. If the null and alternative hypotheses are H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$ then the test for testing the null hypothesis is right-tailed because the alternative hypothesis is right-tailed. Similarly, if the null and alternative hypotheses are H_0: $\mu = \mu_0$ H_1: $\mu \neq \mu_0$ then the test for testing the null hypothesis is left-tailed because the alternative hypothesis is left-tailed. The above discussion can be summarised in the Table 1 below:

Table 1: Null and Alternative Hypothesis (Right and Left tailed test)

Null Hypothesis	Alternative Hypothesis	Types of Critical Region
H_0: $\mu = \mu_0$	H_1: $\mu \neq \mu_0$	Two-tailed test having critical regions under both tails
H_0: $\mu = \mu_0$	H_1: $\mu \neq \mu_0$	Right-tailed test having critical region under right tail only
H_0: $\mu = \mu_0$	H_1: $\mu \neq \mu_0$	Left-tailed test having critical region under left tail only

9.2.3 Errors in Hypothesis

If the value of the test statistic falls in rejection (critical) region then we reject the null hypothesis and if it falls in the non-rejection region then we do not reject the null hypothesis. A test statistic is calculated based on 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 sub-standard when it is not. It is an error caused by to poor or inappropriate (faulty) sample. Similarly, a packet of screws may infer good when it is sub-standard. So, we can commit two kinds of errors while testing a hypothesis which are summarised in Table 2 which is given below:

Table 2: Types of Error

Decision	**H_0 True**	**H_1 True**
Reject H_0	Type I Error (α)	Correct decision
Do not Reject H_0	Correct Decision	Type II Error (Power of test) β

Let us take a situation where a patient suffering from high fever reaches a doctor. Suppose the doctor formulates the null and alternative hypotheses as

H_0: The patient has a Stomach Infection

H_1: The patient has not a Stomach Infection

The following cases arise:

Case I: Suppose that hypothesis H_0 is true, that is, the patient is a Stomach Infection and after observation, pathological and clinical examination, the doctor rejects H0, that is, he/she declares him/her a non-Stomach Infection patient. It is not a correct decision and he/she commits an error in a decision known as a type-I error.

Case II: Suppose that hypothesis H_0 is false, that is, the patient is a non-Stomach Infection patient and after observation, the doctor rejects H_0, that is, he/she declares him/her a non-Stomach Infection patient. It is a correct decision.

Case III: Suppose that hypothesis H_0 is true, that is, the patient is a Stomach Infection patient and after observation, the doctor does not reject H_0, that is, he/she declares him/her a Stomach Infection patient. It is a correct decision.

Case IV: Suppose that hypothesis H_0 is false, that is, the patient is a non-Stomach Infection patient and after observation, the doctor does not reject H_0, that is, he/she declares him/her a Stomach Infection patient. It is not a correct decision and he/she commits an error in a decision known as a type-II error.

9.2.4 LEVEL OF SIGNIFICANCE

The level of significance is the probability of rejecting a true null hypothesis that is the probability of Type I error and is denoted by α. The frequently used values of α are 0.05 (i.e., 5 %); 0.01(i.e., 1 %); 0.1(i.e., 10 %), etc. When α = 0.05 it means that the level of significance is 5%, α = 0.01 which means a 1% level of significance. α = 0.01 which means 10% level of significance. In fact, α specifies the critical region. If the calculated value of the test statistic lies in the rejection (critical) region, then we reject the null hypothesis and if it lies in the 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.

9.2.5 Confidence Interval

Confidence interval is the interval marked by limits within which the population value lies by chance and the hypothesis is considered to be acceptable. If an observed value falls in the confidence interval H_0 is accepted.

9.2.6 Degree of Freedom

Degree of freedom refers to the number of values that are free to vary after we have given the number of restrictions imposed upon the data. It is commonly abbreviated by df. In statistics, it is the number of values in a study that are free to vary. The statistical formula to find out how many degrees of freedom are there is quite simple. It implies that degrees of freedom are equivalent to the number of values in a data set minus 1, and appears like this:

$df=N-1$

Where N represents the number of values in the data set (sample size).

That being said, let’s have a look at the sample calculation.

If there is a data set of 6, (N=6).

Call the data set X and make a list with the values for each data.

For this example, data, set X of the sample size includes: 10, 30, 15, 25, 45, and 55

This data set has a mean, or average of 30. Find out the mean by adding the values and dividing by N:

$(10 + 30 + 15 + 25 + 45 + 55)/6= 30$

Using the formula, the degrees of freedom will be computed as df = N-1:

In this example, it appears, df = 6-1 = 5

This further implies that, in this data set (sample size), five numbers contain the freedom to vary as long as the mean remains 30.

9.2.7 Power of Test

Nowadays use of p-value is becoming more and more popular because of the following two reasons:

- most statistical software provides a p-value rather than a critical value.
- p-value provides more information compared to critical value as far as rejection or not rejection of H_0

Moving in this direction, we note that in scientific applications one is not only interested in rejecting or not rejecting the null hypothesis but he/she is also interested in assessing how strong the data has the evidence to reject H_0.

This smallest level of significance (α) is known as the "p-value". The p-value is the smallest value of the level of significance(α) at which a null hypothesis can be rejected using the obtained value of the test statistic. The p-value is the probability of obtaining a test statistic equal to or more extreme (in the direction of sporting H_1) than the actual value obtained when null hypothesis is true.

9.3 CRITICAL/ ACCEPTANCE REGION

Results from statistical tests will fall into one of two regions: the rejection region, which will lead you to reject the null hypothesis, or the acceptance region**,** where you provisionally accept the null hypothesis. The acceptance region is the complement of the rejection region; If your result does not fall into the rejection region, it must fall into the acceptance region.

Critical values are values separating the values that support or reject the null hypothesis and are calculated based on alpha. Based on the alternative hypothesis, three cases of critical region arise:

Case A) Two-tailed test:

In this hypothesis testing method, the critical region lies on both sides of the sampling distribution. It is also known as a non - non-directional hypothesis testing method. The two-tailed test is used when it needs to be determined if the population parameter is assumed to be different than some value. The hypotheses can be set up as follows:

H_0: the population parameter = some value

H_1: the population parameter ≠ has some value

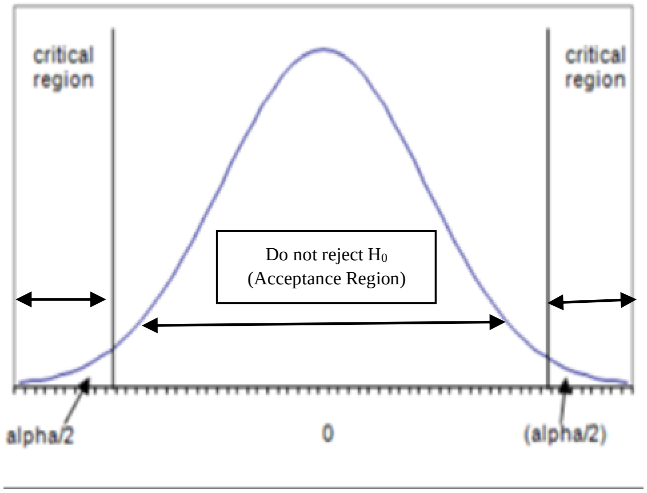
The null hypothesis is rejected if the test statistic has a value that is not equal to the critical value.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$

That interval is the acceptance region. Figure 1 shows a two-tailed having 95 % acceptance region i.e., 2.5 % from the left side and 2.5 % from the right side.

Figure1: Two-tailed hypothesis testing



Case B) Left-tailed test:

The left tail test is also known as the lower tail test. It is used to check whether the population parameter is less than some value. The hypotheses for this hypothesis testing can be written as follows:

H_0: The population parameter is ≥ some value

H_1: The population parameter is < some value.

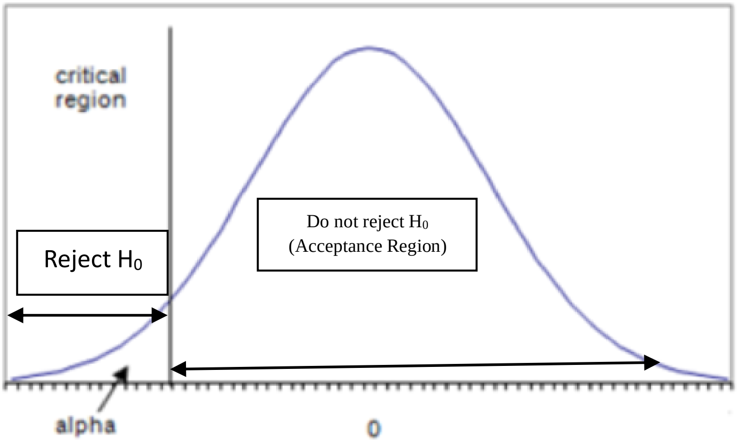
The null hypothesis is rejected if the test statistic has a value lesser than the critical value.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu < \mu_0$

It also shows that 95 % of the right region is an acceptance region.

Figure 2: Left-tailed hypothesis testing



Case C) Right-tailed test:

The right tail test is also known as the upper tail test. This test is used to check whether the population parameter is greater than some value. The null and alternative hypotheses for this test are given as follows:

H0: The population parameter is ≤ some value

H1: The population parameter is > some value.

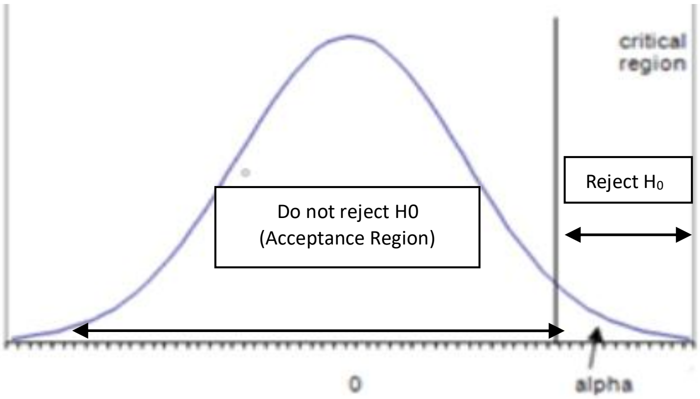
If the test statistic has a greater value than the critical value then the null hypothesis is rejected.

Therefore, H_0: $\mu = \mu_0$

H_1: $\mu > \mu_0$

It also shows 95 % of the left region as an acceptance region

Figure 3: Right-tailed hypothesis testing



9.4 HYPOTHESIS TESTING PROCEDURE

Testing of hypothesis is a huge demanded statistical tool by many disciplines and professionals. It is a step-by-step procedure as you will see in the next three units through a large number of examples. The following steps are involved in hypothesis testing:

Step I: First of all, we have to set up null hypothesis H_0 and alternative hypothesis H_1. Suppose, we want to test the hypothetical / claimed / Testing of Hypothesis assumed value μ_0 of parameter μ.

So, we can take the null and alternative hypotheses as H_0: $\mu = \mu_0$

H_1: $\mu \neq \mu_0$ (for the two-tailed test)

While one- tail test as:

H_0: $\mu = \mu_0$ and H_1: $\mu > \mu_0$ (Right-tailed)

H_0: $\mu = \mu_0$ and H_1: $\mu < \mu_0$ (Left-tailed)

In case of comparing the same parameter of two populations of interest, say, μ_1 and μ_2, then our null and alternative hypotheses would be

H_0: and $\mu_1 = \mu_2$ and H1: $\mu_1 \neq \mu_2$ (for two-tailed test)

While one- tail test as:

H_0: $\mu_1 \leq \mu_2$ and H_1: $\mu_1 > \mu_2$

H_0: $\mu_1 \geq \mu_2$ and H_1: $\mu_1 < \mu_2$

Step II: After setting the null and alternative hypotheses, we establish a 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. The most common value of α is 0.05 or 5%. Other popular choices are 0.01 (1%) and 0.1 (10%).

Step III: The third step is to choose an appropriate test statistics form like Z (standard normal), $\chi 2$, t, F or any other well-known in the literature.

Step IV: Obtain the critical value(s) in the sampling distribution of the test statistic and construct the rejection (critical) region of size α. Generally, critical values for various levels of significance are put in the form of a table for various standard sampling distributions of test statistics such as Z-table, χ 2 -table, t-table, etc.

Step V: After that, compare the calculated value of test statistic obtained from Step IV, with the critical value(s) obtained in Step V and locate the position of the calculated test statistic, that is, it lies in the rejection region or non-rejection region.

Step VI: ultimately testing the hypothesis, we have to conclude.

It is done as explained below:

(i) If the calculated test statistic value lies in the rejection region at the significance level, then we reject the null hypothesis. It means that the sample data provide us with sufficient evidence against the null hypothesis and there is a significant difference between hypothesized value and observed value of the parameter.

(ii) If the calculated test statistic value lies in the non-rejection region at the significance level, then we do not reject the null hypothesis. It means that the sample data fails to provide sufficient evidence against the null hypothesis and the difference between hypothesized value and observed value of the parameter due to sample fluctuation.

Nowadays the decision about the null hypothesis is taken with the help of p-value. The concept of p-value is very important because computer packages and statistical software such as SPSS, STATA, MINITAB, EXCEL, etc., all provide p-value.

Example 1: Mean average weight of men is greater than 100kgs with a standard deviation of 15kgs. 30 men are chosen with an average weight of 112.5 Kg. Using hypothesis testing, check if there is enough evidence to support the researcher's claim. Check the significance at the 5 % level.

Step 1: This is an example of a right-tailed test. Set up the null hypothesis and alternative hypothesis as

H0: μ = 100.

The alternative hypothesis is given by

H1: μ > 100.

Step 2: Level of Significance:

As this is a one-tailed test,

α = 5%. This can be used to determine the critical value.

1 - α = 1 - 0.05 = 0.95

0.95 gives the required area under the curve. Now using a normal distribution table, the area 0.95 is at z = 1.645. A similar process can be followed for a t-test. The only additional requirement is to calculate the degrees of freedom given by n - 1.

Step 3. Select the statistic

Here, we must use the z statistic to test the null hypothesis since the variance is known.

Step 4. Find the critical region:

The z-value obtained from the statistical Table for z is 1.645. Hence, the critical region for a one-tailed test is: z > 1.645.

Step 5: Calculate the z-test statistic. This is because the sample size is 30. Furthermore, the sample and population means are known along with the standard deviation.

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

μ = 100, $\bar{X}$= 112.5, n=30, σ= 15

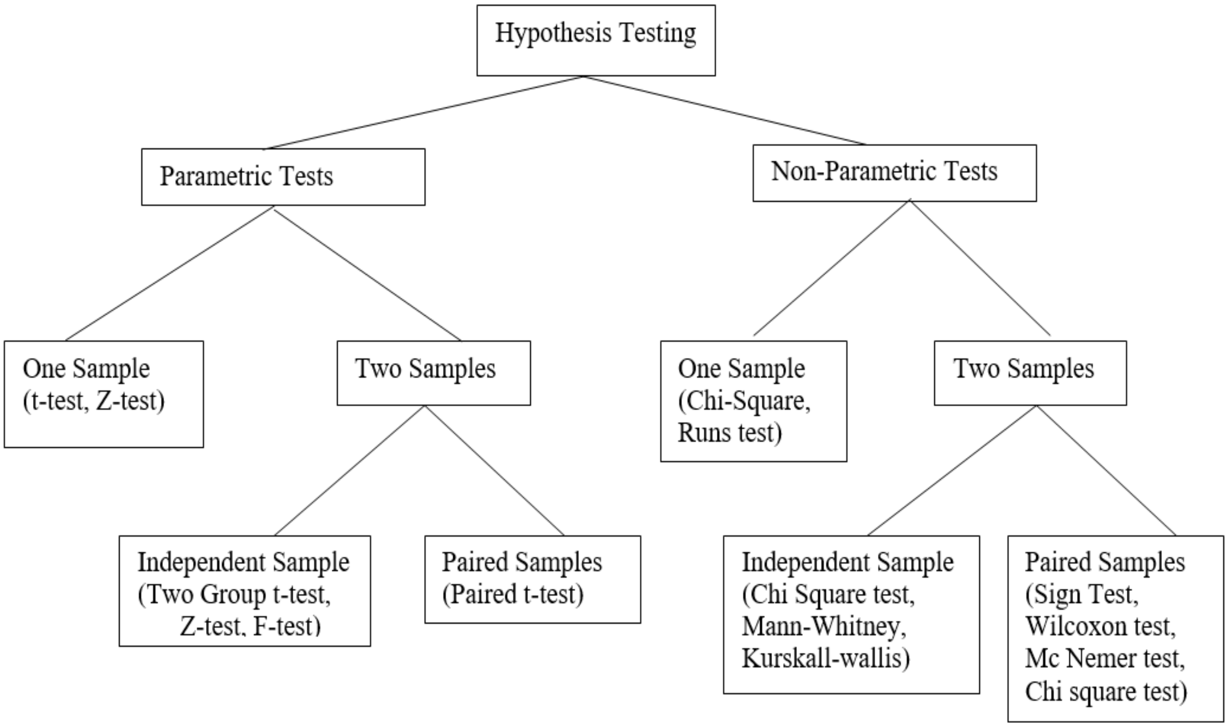
$$Z = \frac{12.5-100}{\frac{15}{\sqrt{30}}} = 4.56$$

Step 6: Conclusion. As Cal Z > tab Z i.e., 4.56 > 1.645 thus, the null hypothesis can be rejected.

9.5 TYPES OF HYPOTHESIS TESTING

Hypothesis testing is a fundamental statistical technique used to make inferences about populations based on sample data. Several types of hypothesis tests are designed for different scenarios and research questions.

- Parametric Tests
- Non-Parametric Tests


Hypothesis Testing
Parametric Tests
Non-Parametric Tests
One Sample (t-test, Z-test)
Two Samples
One Sample (Chi-Square, Runs test)
Two Samples
Independent Sample (Two Group t-test, Z-test, F-test)
Paired Samples (Paired t-test)
Independent Sample (Chi Square test, Mann-Whitney, Kurskall-wallis)
Paired Samples (Sign Test, Wilcoxon test, Mc Nemer test, Chi square test)

9.5.1 PARAMETRIC TEST

These tests are based on several assumptions about the parent population from which the sample was taken. The assumptions may relate to sample size, distribution type, or population characteristics like mean, standard deviation, etc. The most widely used parametric tests are the Z-test, t-test, and $\chi 2$ test (although x is considered a nonparametric test when used as a test of independence or good of fit). Since they use interval and ratio data, parametric tests are more potent than nonparametric tests.

Parametric tests are based on certain assumptions.

The observations being tested should be independent so that the inclusion of one set of observations does not affect the subsequent observations,

- normality of distribution
- It requires interval or ratio measurement scales so that arithmetic operations can be applied to them

The "Z" test is used for t-distribution and binomial or Poisson distribution also when the sample size is very large on the presumption that such a distribution tends to approximate normal distribution as the sample size becomes larger. This Z value is compared with the calculated Z-statistic for judging the significance of the measure concerned. The 't' test is a univariate test that uses t-distribution for testing sample mean and proportion when the size of sample is small (i.e., less than 30). The t-distribution is a symmetrical bell-shaped curve. The variance of t-distribution approaches the variance of the standard normal distribution as the sample size increases. Hence the widely practiced rule of thumb is that n > 30 is considered large and for such sample size normal distribution is used and for n < 30 t-distribution is used.

'F'-test is based on F-distribution. It is generally used to compare the variance of two sets of observations. F-distribution is a frequency distribution that uses two sets of degrees of freedom i.e., one in numerator and one in denominator. ANOVA is a case of using F-test to compare variance. Chi-square considered as a parametric test is used to compare a sample variance to some theoretical population variance. It is based on chi-square distribution.

9.5.2 NON-PARAMETRIC TEST

Non-parametric tests, also known as distribution-free tests, are a category of statistical tests that do not make strong assumptions about the underlying distribution of the data. These tests are used when the data do not meet the assumptions of parametric tests, which assume that the data follow a specific distribution, such as a normal distribution. Non-parametric tests are often used when dealing with ordinal or nominal data, small sample sizes, or data that are not normally distributed.

Here are some common non-parametric tests:

Mann-Whitney U Test (Wilcoxon Rank-Sum Test): This test is used to compare two independent groups to determine if there is a significant difference between them. It is used as a non-parametric alternative to the independent samples t-test.

Wilcoxon Signed-Rank Test: This test is used to compare two related (paired) groups when the data is not normally distributed. It is an alternative to the paired samples t-test.

Kruskal-Wallis Test: This is a non-parametric alternative to the one-way ANOVA test. It is used to compare three or more independent groups to determine if there are significant differences between them.

Friedman Test: This is the non-parametric counterpart of the repeated measures ANOVA. It is used to test for differences between multiple related groups when the data are not normally distributed.

Chi-Square Test: The chi-square test is used to analyse the association between categorical variables. It can be used for tests of independence (chi-square test of independence) or tests of goodness-of-fit (chi-square goodness-of-fit test).

Mann-Whitney-Wilcoxon Test (MWW): This is an extension of the Mann-Whitney U test for comparing more than two independent groups.

Sign Test: A non-parametric test for comparing paired data. It tests whether the median of the differences between paired observations is significantly different from zero.

Runs Test: This test is used to determine whether a sequence of data points is randomly ordered or exhibits some systematic pattern.

Non-parametric tests are valuable tools in statistics when assumptions of normality or other parametric assumptions are not met. They are robust and provide a way to perform statistical analysis when the data does not conform to the assumptions of parametric tests. However, they are generally less powerful than their parametric counterparts when the assumptions of parametric tests are met, so it's important to choose the appropriate test based on the nature of your data and research questions.

9.6 CHI-SQUARE TEST

The Chi-Square Test is based on chi-square distribution and is a nonparametric test (or distribution-free test) as it does not require any assumptions for population parameters. This test helps to determine the difference between observed and expected data. The main objective of the Chi-square test is to find out whether a difference between given categorical (qualitative) variables is due to chance or any link between them. As categorical variables can be nominal or ordinal, having few particular values so cannot be expressed with the help of normal distribution. For example: a tea-selling firm wants to find out the relationship between consumer's gender, location and flavour of tea. Here difference between two categorical variables can be due to chance or because of some specific relationship.

Conditions for the validity of the Chi-square test:

- Sample observations should be independent which means all individual items are included only once in the sample.
- The cell frequencies should be linear only i.e., ∑O = ∑ E = N.
- N, the total frequency should be reasonably large (greater than 50).
- Theoretical frequency should not be less than 5, if so, then use the pooling technique (adding preceding or succeeding frequency or frequencies in theoretical frequency which is less than 5) and accordingly adjust degrees of freedom.
- The data should be given in original units only and not in relative or proportion form.

9.7 APPLICATIONS

Chi-square tests are commonly used to test null hypothesis related to the size of inconsistency between the expected results and actual results by using the degree of freedom. There are the following common Chi-Square tests which we will discuss in detail:

1. $squared\ test$ test of the goodness of fit.
2. $squared\ test$ test for the independence of attributes.
3. χ^2 test if the population has a specified value of the variance σ^2.
4. $squared\ test$ test of equality of several population proportions.

9.7.1 Chi-Square Test of Goodness of Fit

This test was first developed by Karl Pearson in 1900 to check the significant differences between experimental values and the theoretical values obtained under some theory or hypothesis. It helps to find out whether the difference between observation and theory is because of fluctuations of sampling or because of the inadequacy of theory to fit the observed data. We can decide whether data values are a good fit for our idea (theory) or whether sample data values represent the entire population.

Steps to compusquared value for drawing a conclusion:

1. Set Null Hypothesis as follows: There is no significant difference between theory and experiment.

   Or

There is no significant difference in observed (experimental) and theoretical (or hypothetical) values.

2. Calculate the expected frequencies (E_1, E_2,….E_n) corresponding to given observed frequencies (O_1, O_2,…..O_n).
3. Calculate the difference between each observed and expected frequency and then square them i.e., $(O – E)^2$.
4. Divide each square of the difference of observed and expected frequency obtained in step 3 by the corresponding expected frequency i.e., $(O – E)^2 / E$.
5. Add the values calculated in step 4 to get $\boldsymbol{\chi^2} = \sum \mathbf{[}(O – E)^2 / E]$
6. Here null hypothesis follows a chi-square distribution with v = (n-1) degrees of freedom (d.f)
7. Check critical (tabulated) values from a table of chi-square distribution for $chi - squared$ for (n-1) d.f at a certain level of significance.
8. Compare calculated and critical values of superscriptof χ^2. If the calculated value is greater than the tabulated value then it is significant and we reject null hypothesis which also means there is a significant difference between the experimental and theoretical values. In other words, differences between observed and expected frequencies cannot be because of fluctuations of sampling.

Example 2: The number of accidents per month in town A given as follows:

Month:	1	2	3	4	5	6	7	8	9	10
No. of accidents:	4	9	6	15	10	14	2	20	8	12

Test the null hypothesis that accident conditions were same during all given months.

Solution: Set Null Hypothesis (H_0): accident conditions were same during all the given months
Or

There is no significant difference between accident conditions during the given months.

Here total number of accidents during 10 months is 100 (4+9+6+15+10+14+2+20+8+12)

So Expected number of accidents (E) = 100/10 = 10

Table 9.1: Calculation o Chi-Square (χ^2)

Month	Observed no. of accidents (O)	Expected no. of accidents (E)	(O – E)	$(O – E)^2$	$(O – E)^2$ / E
1	4	10	-6	36	3.6
2	9	10	-1	1	0.1
3	6	10	-4	16	1.6
4	15	10	-5	25	2.5
5	10	10	0	0	0
6	14	10	4	16	1.6
7	2	10	-8	64	6.4
8	20	10	10	100	10
9	8	10	-2	4	0.4
10	12	10	2	4	0.4
Total	100	100			26.6

$\chi^2 = \sum [(O – E)^2 / E] = 26.6$ (calculated value)

d.f. = n-1= 10 – 1= 9 d.f. at 5% level of significance = 16.919 (tabulated value)

Here calculated value is greater than the tabulated value, it is significant and null hypothesis is rejected. Hence it is concluded that accident conditions were not the same in given 10-month period.

9.7.2 Chi-Square test for independence of attributes

The dictionary meaning of attributes is quality or characteristic for example health, employment, honesty, beauty, gender etc. Now suppose there are two attributes A and B, divided into m and n classes respectively, such that $A_{1,}$ A_2,….,A_m classes concerning the attribute A and B_1, B_2,….., B_n classes with respect to attribute B. This type of classification is also known as manifold classification. The frequency distribution of A and B attributes can be expressed in following m × n manifold contingency table.

Table 9.2: m × n Manifold Contingency Table

Attributes	**B_1**	**B_2**	**…**	**Bj**	**…**	**B_n**	**Total**

A_1	**$(A_1\ B_1)$**	**$(A_1\ B_2)$**	**…**	**$(A_1\ B_j)$**	**…**	**$(A_1\ B_n)$**	**(A_1)**
A_2	**$(A_2\ B_1)$**	**$(A_2\ B_2)$**	**…**	**$(A_2\ B_j)$**	**…**	**$(A_2\ B_n)$**	**(A_2)**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
Ai	**$(A_i\ B_1)$**	**$(A_i\ B_2)$**	**…**	**$(A_i\ B_j)$**	**…**	**$(A_i\ B_n)$**	**(A_i)**
…	**…**	**…**	**…**	**…**	**…**	**…**	**…**
A_m	**$(A_m\ B_1)$**	**$(A_m\ B_2)$**	**…**	**$(A_m\ B_j)$**	**…**	**$(A_m\ B_n)$**	**(A_m)**
Total	**(B_1)**	**(B_2)**	**…**	**(B_j)**	**…**	**(B_n)**	**N**

In the above table given population consisting of N items is divided into m mutually exclusive and exhaustive classes $A_{1,}$ A_2,….,A_m of attribute A and n mutually exclusive and exhaustive classes B_1, B_2,…..,B_n of attribute B.

9.7.2.1 2*2 Contingency Table

As we have discussed above manifold contingency classification of attributes, similarly we can have two attributes A and B, divided into 2 classes each, such that A_1 and A_2 classes with respect to the attribute A and B_1 and B_2 classes with respect to attribute B. This type of classification is also known as the 2 × 2 classification. The frequency distribution of A and B attributes can be expressed in following 2 × 2 contingency table. Here population consisting of N items is divided into 2 mutually exclusive and exhaustive classes A_1 and A_2 of attribute A and 2 mutually exclusive and exhaustive classes B_1, and B_2 of attribute B.

Table 9.3: 2 × 2 Contingency Table

Attributes	**B_1**	**B_2**	**Total**
A_1	**$(A_1\ B_1)$**	**$(A_1\ B_2)$**	**(A_1)**
A_2	**$(A_2\ B_1)$**	**$(A_2\ B_2)$**	**(A_2)**
Total	**(B_1)**	**(B_2)**	**N**

Steps to compute χ^2 value for drawing the conclusion

1. Set Null Hypothesis as follows

   The two attributes are independent

2. Calculate the expected frequencies (E) corresponding to all observed frequencies (O) for example: for $(A_1\ B_2) = \frac{(A1)\ (B2)}{N}$.
3. Calculate $(O – E)^2$ / E.

4. Add the values calculated in step 3 to get $\chi^2 = \sum [(O - E)^2 / E]$
5. Compare the calculated value of $\boldsymbol{\chi^2}$ with its tabulated value at a certain significance level for (no. of rows-1) (no. of columns -1) = (2-1) (2-1) = 1 d.f and draw the conclusion.

Example 3: From the following table test whether the colour of the sons' eyes is associated with that of the fathers'

Father eye colour	Son eye colour		Total
	Not Black	Black	
Not Black	230	148	378
Black	151	471	622
Total	381	619	1000

Solution: Set null hypothesis, i.e., two attributes are independent.

Or attributes father's eye colour and son's eye colour are independent.

Now calculate expected frequencies corresponding to all observed frequencies. The expected frequency for 230 can be written as E (230) = $\frac{378 \times 381}{1000}$ = 144.018

Similarly, E (148) = $\frac{378 \times 619}{1000}$ = 233.982, E (151) = $\frac{622 \times 381}{1000}$ = 236.982,

E (471) = $\frac{622 \times 619}{1000}$ = 385.018

Table 9.4: Calculation Of $\boldsymbol{\chi^2}$

O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
230	144.081	85.919	7382.0745	51.2355
148	233.982	-85.982	7392.9043	31.5960
151	236.982	-85.982	7392.9043	31.1960
471	385.081	85.919	7382.0745	19.1701
1000	1000	0		133.1976

$\boldsymbol{\chi^2} = \sum [(O - E)^2 / E]$ =133.1976 (calculated value)

d.f. = (2-1) (2-1) = 1d.f. at 5% level of significance = 3.841 (tabulated value)

Here calculated value is greater than tabulated value so it is highly significant and we reject the null hypothesis. In other words, colour of sons' eyes is associated with that of fathers' eyes.

9.7.2.2 Yates Correction

If in the 2 × 2 table any cell frequency is less than 5 then for application of chi-square pooling technique will not be useful as it will lead to a loss of degree of freedom. After adjustment with the help of the pooling technique to make cell frequency greater than 5, d,f will be zero only. So here we use Yates Correction for 'continuity', under which 0.5 is added to the cell frequency which is less than 5 and adjusting remaining cell frequencies so that totals remain same.

Suppose we have following 2 × 2 table

Attributes	B_1	B_2	Total
A_1	2	10	12
A_2	6	6	12
Total	8	16	24

In the above table A_1B_1 cell frequency is less than 5 so as per Yates correction it can be adjusted as follows:

Attributes	B_1	B_2	Total
A_1	2.5	9.5	12
A_2	5.5	6.5	12
Total	8	16	24

The rest of the procedure is same for calculating chi-square.

9.7.3 Chi-Square test if the population has a specified value of the variance σ^2

Under this, we can test if the given population has a specified variance ($\sigma^2 = \sigma_0^2$). Now suppose we have a random sample (x_1, x_2, x_3,…..x_n) of size n from the given population then to test about the specified value of population variance following steps are given

1. Set null hypothesis

   H_0: $\sigma^2 = \sigma_0^2$

2. calculated value of $\boldsymbol{\chi^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma_0^2} = \frac{\mathrm{ns}^2}{\sigma_0^2}$ which follows chi square distribution with (n-1) d.f where $s^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x_i - \bar{x}})^2$**,** is a sample variance.

3. Find out tabulated value at (n-1) d.f at certain level of significance.

4. Compare calculated and tabulated values to draw conclusion.

Example 4: A sample of 15 values shows the standard deviation to be 6.4. Does this agree with the hypothesis that the population standard deviation is 5, the population being normal?

Solution: set null hypothesis (H_0): population standard deviation is 5.

Here $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{15 \times 40.96}{5 \times 5}$ = 24.576 (calculated value)

At (15-1) d.f for 5% level of significance, tabulated value of chi-square is 23.685

Calculated value is greater than tabulated value so null hypothesis is rejected. In other words, population s.d is not 5

9.7.4 Chi Square Test of Equality of Several Population Proportions

The Z test of equality of two population proportions for large samples is extended to Chi-Square test for several population mean. Here we take independent random samples of pre-determined size from several populations and write down the population proportions (i.e., $P_1 = P_2 = P_3 = P_4 = \ldots.= P_n$) into two categories (for example: married and unmarried). The null hypothesis is $P_1 = P_2 = P_3 = P_4 = \ldots.= P_n$, which is the chi square statistic for testing the independence of two variables for 2 × n contingency table.

$$\chi^2 = \sum [(O - E)^2 / E]$$

d.f = (r-1) (n-1) = (2-1) (n-1), where n is no. of columns

Expected frequency will be calculated in same way as we have calculated in earlier tests i.e.,

$$\frac{\text{Row Total} \times \text{Coloumn Total}}{\text{Grand total}}$$

Example 5: In a survey, it is found that 77 of 220 housewives in city A, 260 of 650 housewives in City B, 72 of 225 housewives in City C, and 120 of 315 housewives in city D watch a daytime popular T.V serial. At 5% level of significance, test if there is no difference between the true proportions of housewives who watch the TV serial in these cities.

Solution: Firstly, write down the information in the table as follows

Serial Watch/Cities	A	B	C	D	Total
House wives watching	77	260	72	120	529

House wives not watching	143	390	153	195	881
Total	220	650	225	315	1410

Null Hypothesis (H_0): all population proportions are equal or $P_1 = P_2 = P_3 = P_4$

Alternative Hypothesis (H_1): all population proportions are not equal or P_1, P_2, P_3, P_4 are not equal.

Now calculate expected frequencies for all observed frequencies

E (77) = $\frac{529 \times 220}{1410}$ = 82.539, E (260) = $\frac{529 \times 650}{1410}$ = 243.865, E (72) = $\frac{529 \times 225}{1410}$ = 84.414,

E (120) = $\frac{529 \times 315}{1410}$ = 118.180, E (143) = $\frac{881 \times 220}{1410}$ = 137.460, E (390) = $\frac{881 \times 650}{1410}$ = 406.134,

E (153) = $\frac{881 \times 225}{1410}$ = 140.585 E (195) = $\frac{881 \times 315}{1410}$ = 196.819

Calculation of Chi-Square

O	E	O - E	$(O – E)^2$	$(O – E)^2/E$
77	82.539	-5.539	30.680	0.371
260	243.865	16.135	260.338	1.067
72	84.414	-12.414	154.107	1.825
120	118.180	1.82	3.312	0.028
143	137.460	5.54	30.691	0.223
390	406.134	-16.134	260.305	0.640
153	140.585	12.415	154.132	1.096
195	196.819	-1.819	3.308	0.016
1410 (total)	1410 (total)			5.266 (total)

$\chi^2 = \sum [(O – E)^2 / E] = 5.266$ (calculated value)

d.f = (r-1) (n-1) = (2-1) (4-1) = 3 d.f

tabulated value at 3 d.f at 5% level of significance = 7.815

here calculated value is less than tabulated value so H_0 is not rejected. In other words, all population proportions are equal or there is no difference between the true proportions of housewives who watch the T.V serial in four given cities.

9.8 SUM UP

Testing of hypotheses means to test the assumption validity through null hypothesis. Results from statistical tests will fall into one of two regions: the rejection region and the acceptance region, rejection region leads you to reject the null hypothesis, or the acceptance region, where you provisionally accept the null hypothesis. The acceptance region is "the interval within the sampling distribution of the test statistic that is consistent with the null hypothesis H_0 from hypothesis testing. Parametric Tests assume that the data follows a specific probability distribution, typically the normal distribution. Common parametric tests include t-tests, analysis of variance (ANOVA), and linear regression. Whereas, Non-Parametric tests make fewer assumptions about the data distribution. They are used when the data is not normally distributed or when you have ordinal or categorical data—for example, the Wilcoxon signed-rank test and the Kruskal-Wallis test. Chi Square Test is based on chi square distribution, which is non parametric test helps to determine the difference between observed and expected data categorical (qualitative) variables.

9.9 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. Define types of hypotheses

Q2. Explain the types of Error

Q3. Define level of significance

Q4. Explain the concept of degree of freedom

Q5. What do you mean by power of a test

Q6. Explain acceptance region

Q7. Discuss rejection region

Q8. What is the meaning of parametric test

Q9. Explain Non-parametric test

B. Long Answer Type Questions

Q1. What do you mean by Chi-Square test? Give the conditions and applications of Chi Square test.

Q2. Discuss the steps of applying Chi-Square tests of Goodness of Fit and Independence of Attributes.

Q3. State the conditions of validity of the Chi-Square test. Also, discuss the applications with their null hypothesis and degrees of freedom of chi-square?

Q4. In a set of random numbers, the digits 0 to 9 were found to have the following frequencies:

Digits:	0	1	2	3	4	5	6	7	8	9
Freq:	43	32	38	27	38	52	36	31	39	24

Test whether they are significantly different from those expected on the hypothesis of uniform distribution. (Calculated Chi Square = 16.33, tabulated for 9 d.f. at 5% l.o.s = 16.92)

Q5. Out of a sample of 120 persons in a village, 76 persons were administered a new drug for preventing corona and out of them 24 persons were attacked by corona. Out of those who were not administered the new drug, 12 persons were not affected by corona:

(a) Prepare 2 × 2 table showing actual and expected frequencies.

(b) Use the Chi-Square test to find out whether the new drug is effective or not.

(Calculated Chi Square = 18.968, tabulated at 5% l.o.s for 1 d.f. = 3.84)

Q10. Discuss the steps of applying Chi-Square tests if population has a specified value of the variance σ^2 and Equality of Several Population Proportions?

<u>9.10</u> <u>SUGGESTED READINGS</u>

- Anderson, D.R.; Sweeney, D.J. and Williams, T.A., "Statistics for Business and Economics", 2nd edition (2011), Thompson, New Delhi.
- Gupta SC: Fundamental of statistics, S. Chand & Company. New Delhi
- Gupta, SP: Statistical Methods, S. Chand & Company. New Delhi
- Kothari, C. R., "Research Methodology", 2nd Edition (2008), New Age International.
- Meyer, P.L. (1990): Introductory Probability and Statistical Applications, Oxford & IBH Pub.
- Monga, GS: Mathematics and Statistics for Economics, Vikas Publishing House, New Delhi.
- Rohatgi, V. K. and Saleh, A.K.M.E. (2010): An Introduction to Probability Theory and Mathematical Statistics, Wiley Eastern.

M.COM

SEMESTER-III

RESEARCH METHODOLOGY AND STATISTICAL ANALYSIS

UNIT 10: INTERPRETATION AND REPORTING: INTERPRETATION OF STATISTICAL DATA REPORT WRITING

STRUCTURE

10.0 OBJECTIVES

After reading this unit, learners can able to know about:

- Importance and precautions of interpretation

- Research report writing
- Qualities of a good research report
- Significance of report writing
- Types of report
- Referencing styles
- Evaluation of the research report

10.1 INTRODUCTION

Interpretation and reporting are important processes in various fields, including research, data analysis, business, and more. They involve analysing data, information, or results and presenting them in a clear and meaningful way to make informed decisions, convey findings, or communicate a message. In other words, Data Interpretation is the process of making sense of numerical data that has been collected, analysed, and presented. It is the process of attaching meaning to the data and establishing meaning out of the data. Interpretation demands fair and careful judgments as it reflects the theoretical and analytical ability of the researcher to penetrate into the data and identify the variables exhibiting a relationship.

Definition: "Interpretation refers to the process of making sense of numerical data that has been collected, analysed and presented".

An example of a social media platform generates vast amounts of data every second, and businesses can use this data to analyse customer behavior, track sentiment, and identify trends. Data interpretation in social media analytics involves analysing data in real-time to identify patterns and trends that can help businesses make informed decisions about marketing strategies and customer engagement.

10.2 BENEFITS/ IMPORTANCE OF INTERPRETATION

- Interpretation plays a crucial role in highlighting the practical applications of research findings and opening them up for utilisation. Utilising the results in practical settings makes the research worthwhile. For the following reasons, interpretation is crucial:
- It is a necessary and significant step at the end of research projects because it allows the researcher to explain the general idea that underlies his findings.

- Research continuity is aided by interpretation. By utilising the same basic principles, the results of this study can be connected to those of previous studies, offering an updated and more insightful understanding of preexisting notions.
- Exploratory interpretation frequently acts as a roadmap for developing hypotheses for framing hypothesis for successive research efforts. It provides a base for carrying forward the research efforts.
- Interpretation gives the researcher the chance to better understand the outcomes of his research endeavour while also enabling him to communicate the findings to others in a more relevant way.
- It is common for interpretation to result in the identification of novel connections and conceptual frameworks. Thus, it creates more opportunities for intellectual pursuits.

10.3 PRECAUTIONS IN INTERPRETATION

While interpreting the results the researcher should be alert and take certain precautions like:

1. Firstly, the researcher should ensure that the data available for interpretation has been collected through reliable sources and is valid. It should be adequate enough to render itself to meaningful interpretation.
2. The statistical methods used for data analysis must be looked into and the researcher should be satisfied that the appropriate methods have been used. An expert's help should be sought to check the reliability and appropriateness of statistical methods since the output of statistical analysis will serve as an input to interpretation.
3. The sampling and non-sampling errors that have crept into the research process must be listed. Generalizations made on a very small sample size or errors in the editing, coding, or tabulating process can severely damage the authenticity of data. Such errors would lead to errors of interpretation and which if committed would nullify the findings of even the best research.
4. A researcher should have a clear understanding of the problem. An inadequate grasp of the broader perspective of the problem and focusing all attention on the immediate aspect may lead to a narrow and limited interpretation.

5. At the other extreme broad generalisations should also be avoided. The endeavour should be to make correct interpretation of the observed occurrences within a specified time, place and conditions and at the same time bring out the latent relationships and occurrences in the open.
6. It has been seen that it is easier to interpret results that conform to existing theories. However, when the results are negative or inconclusive then interpretation becomes difficult. In such a situation the entire research process and methodologies should be carefully scrutinized.
7. Interpretation should not be left in the hands of inexperienced people. Objectivity, dexterity, and professionalism are the qualities that a researcher interpreting the results should possess. Further, to the extent possible, interpretation should be done by people who have been associated with the project right from the beginning.
8. Lastly, interpretation should consider dynamism. The data of which interpretation has been done relates to a single point in time or period of time. in the past and by the time the interpretation is done, things might have changed. Hence interpretation should be done keeping in mind the past, as well as the changed conditions.

10.4 REPORT WRITING

Report writing is most crucial as it is through the report that the findings of the study and their implications are communicated to the readers. As a matter of fact, even the most brilliant hypothesis, highly well-designed and conducted research study and the most striking generalizations and findings are of little value unless they are effectively communicated to readers. Most people will not be aware of the amount and quality of work that has gone into the study, while much hard work and care might have been put in at every stage of the research, what all readers see is the report. Therefore, the whole purpose of working so hard is defeated if appropriate justice is not done in writing the report. The very purpose of writing the report is to be able to communicate to the readers the nature of methodology followed in doing research and in deriving the findings at which one has arrived.

Definition: A research Report is a "Systematic, articulate and orderly presentation of research work in a written form”.

10.5 QUALITIES OF A GOOD RESEARCH REPORT

A research report is documentary evidence of the research effort. It serves as a guide for the authorities to evaluate the quality of the entire research effort and decisions are guided by the results presented in the report. Hence a research report should possess certain basic qualities as discussed below:

- **Communicate with the reader**: A report is of some worth only if its contents are easily understood by the readers. The report should be specific to port the readers in terms of details mentioned, presentation and technical language used. "The readers of your report are busy people, and very few of them can balance a research report, a cup of coffee, and a dictionary at one time". Technical terms of unavoidable, then should be explained clearly in the glossary.
- **Completeness**: A report is considered complete if it provides all the relevant information and answers the problem adequately. It is not the length but the content which determines the completeness. Too short a report may omit necessary information whereas the same may be lost in the cluster of a lengthy report. The length of the report is generally determined by the characteristics of the reader.
- **Brief**: To be concise is to express a thought completely and clearly in the fewest words possible. However, shortness should not be achieved at the expense of completeness. The researcher is unable to find the right phrase or word to express his idea, he tends to write around it, restating it several times hoping that repetition will hide the lack of poor expression.
- **Accurate**: It is possible that despite the input being accurate, the output i.e., the report may develop inaccuracies. These inaccuracies generally result from carelessness in handling data, grammatical errors, concept phrasing, etc. Hence a good report should be free from such inaccuracies.
- **Objective**: Objectivity requires courage to present and defend the results as they actually are, rather than twisting them to suit somebody's preferences.
- **Logical**: A good report should be properly structured and there should be logical. A logically written report means a clear report. The sequence of various sections should be logical and an outline of the major points should be clear. Generally, clarity demands writing and rewriting till the time the ambiguity is removed. Use of headings and subheadings wherever required.

Here think what you want to say. Write your sentence. Then strip it of all adverbs and adjectives.

- **Professional appearance**: quality of the paper print and cover should be good. Standardization should be maintained throughout the report. Tables, graphs, pictures should be added wherever required.

Therefore, the quality of the report is dependent on:

- written communication skills,
- clarity of thoughts,
- ability to express thoughts in a rational and sequential manner,
- knowledge of the subject area.

The use of statistical procedures definitely enhances the quality of work done by the researcher. It also reinforces the validity of one's conclusion and arguments. The use of graphs, tables and diagrams at appropriate places is likely to make the report more attractive and easier to understand for its readers. The most important point to be kept in mind while writing a research report is its intended readers. A report directed to fellow social scientists will be different in certain respects from a report which is meant for laypersons. Whoever is the audience, two general considerations must weigh heavily in the minds of the researchers engaged in writing their research reports:

- What does the audience (or readers) want or need to know about the study?
- How can this information be best presented?

Writing for research is regulated in that you must exercise great caution in what you write, the words you use, the way you present your thoughts, and the reliability and validity of the data you use to support your conclusions. The greatest difference between research writing and other types of writing is the level of intellectual rigour that is necessary. Writings for research projects ought to be totally precise, understandable, rational, crisis-free, and ambiguity-free. Avoid making assumptions about what readers already know about the study. The researcher should be able to provide solid, scientific evidence to back up any claims they make in the report, and their arguments should be convincing to the reader. One needs to avoid superficial language in the report. Even the best researchers make a number of drafts before writing up their final report.

The success will depend on how intelligently it is planned to meet all the objectives it has to fulfil.

A research report should essentially satisfy the following requirements:

- It should include all material of information relevant to the research.
- It should show the information in accordance with a predetermined plan that is based on a classification system and logical analysis of the relevant material.
- It should make this idea so clear that readers may understand it with ease.
- It should be expressed in a clear, uncomplicated manner that eliminates any chance of misunderstanding.
- You should use tables and graphs to supplement it.
- It must include references and appendices.

10.6 SIGNIFICANCE OF REPORT WRITING

- Major component
- Findings are brought into light
- Medium to communicate research work with relevant people.
- Contributes to the body of knowledge
- Effective way of conveying the research work
- Reference material
- Aid for decision making

10.7 TYPES OF REPORTS

- Written Report
- Oral Report

10.7.1 THE WRITTEN REPORT

The most popular type of reports are written ones, which can be classified in a variety of ways. For example, a report's length determines whether it qualifies as a "long report" or a "short report," similar to a progress report. Reports can also be classified as analytical and informational based on their functional purpose. The informational report contains only the facts; it does not analyse or draw any conclusions. The examination report examines the facts, but the analytical report goes one step further and provides recommendations and findings in addition to the analysis. The reports can be divided into three categories based on the relationship between the writer and reader: administrative, professional, and independent. A professional report is prepared by someone from outside the organisation to which it is to be submitted, whereas an administrative report is made by a member of the same organisation. A report that is released for the public's benefit is considered independent. On this basis, a report may be a technical report or a popular report, both of which are long reports.

a) Technical Report: A technical report is a long report containing full documentation. Since everything is described in detail it serves as a source document for future research.

b) Management Report: There are situations when the target audience is more interested in the findings presented in a simple manner rather than the methodology. The approach encourages rapid reading, a quick understanding of the findings and a proper grasp of the implications of the study. A management report should stress on the following things:

   - Objectives should be simple with clear.
   - Pictures and graphs should be used.
   - Findings should be clear.
   - The paragraphs should be short and direct.
   - Important points should be highlighted.

10.7.1.1 STEPS IN WRITING THE REPORT

Writing a report is time-consuming work and goes through numerous drafts before the final presentation is ready. The various steps involved are as follows:

(i) Report Format

The report format can be of three types: logical, Psychological and Chronological

- Logical pattern: It follows an inductive approach where associations are developed between one thing and another by means of analysis.
- Psychological pattern: This approach is the opposite of the logical pattern where the most critical information is stated first and thereafter the findings are stated in a manner that justifies the conclusion.

- Chronological pattern: The reporting of events is done along the time dimension i.e., the events that happened first are stated first and others are stated in the order of their occurrence.

(ii) Preparing a Report Outline

Once the approach to be used to analyse the data has been finalized, the next stage is of preparing an outline. An outline can be prepared according to two styles.

- Topic outline: This uses few keywords, on the assumption that the writer knows its significance and later while writing the detailed report will recall the entire argument.
- Sentence outline: This expresses the essential thoughts in a sentence form and is the preferred method for new inexperienced researchers. Since most of the thoughts are outlined initially and only elaboration and explanation are left for subsequent stages, there is less chance of jumbling and error.

(iii) Preparation of Rough Draft

When the report outline has been decided, it is time to make decisions on the settlement of graphics, tables and charts. The outline prepared in the earlier section is elaborated and the research objective, methodology, findings and recommendations are written out. It is time to write down the account of the entire research process.

(iv) Rewriting and Refining the Report

This stage is a slow and careful step where the writer has to spend more time locating and correcting the errors. Weak points need to be identified in the report format, report outline, grammatical errors in tables and graphs and writing. While rewriting, the writer should ensure comprehensibility, continuity and accuracy. The writer should not submit the report hurriedly but should show patience and attention to each and every detail.

(v) Preparing Bibliography

A bibliography documents the sources used by the researcher in writing the report in an alphabetic order. It contains a list of all the work that the researcher has consulted in the course of his study. In certain situations, the organisation to which the report has to be submitted also states the format of giving a bibliography. Citation, style and format are unique to the purpose of the report.

(vi) Writing the Final Proof

Setting the earlier draught aside for a day or two is a good idea before you start writing the final proof. The researcher can approach the revised draught with a new and critical perspective because to this gap. A clear, objective statement with a high level of consistency and clarity should be the final proof. The final product should be eye-catching enough to entice readers to take it up and captivating enough to hold their interest as they turn the pages. The final proof or report ought to pique readers' intellectual curiosity and broaden their understanding as well as that of the researcher. As said in the previous section, your report needs to embody every quality that makes an excellent report.

10.7.2 ORAL REPORT

An oral report is the presentation of information through spoken word. He/ She has the advantage of creating an interactive environment where they share the information. However, it has the disadvantage that there is no permanent record of the report and the reader does not have the advantage of controlling the pace of presentation. In the case of a written report, if the reader has any doubt or confusion he can refer back to the document and read it as many times and at the place he desires. Hence it becomes important that the presentation of the oral report is of utmost importance.

Key to Effective Oral Presentation

- A written report may not be necessary for an oral report, but the researcher should nonetheless create a thorough description of the format beforehand. After that, the presentation needs to be practiced in front of the audience multiple times before it is given.
- The use of graphs, tables, and diagrams might help to partially compensate for the oral report's shortcoming—namely, the absence of a hard copy. To display the information, the presenter can use audiovisual devices like screens or overhead projectors.
- an essential aspect of an effective presentation is to maintain eye contact with the audience. It makes the targets feel participative and they are more receptive to the information being

presented by the researcher. The presentation can be made more interesting with the help of examples, stories and quotations wherever appropriate.

- Another point to remember is the body language. Each stance and gesture of the body conveys a specific meaning e.g., emphatic gestures are used to emphasize a point, and encouraging gestures can be used to elicit a response from the audience. The volume, pitch, and tone all contribute to the effectiveness of the presentation.
- While delivering the report the two extremes, memorization and reading should be avoided. Memorization creates a speaker-centered approach whereas reading makes the presentation dull and listless. The presentation should involve brief notes and while delivering the way of expressing, evolves naturally and the presenter develops phrases in a conversational manner. This approach is audience-centered and more effective.

10.8 REFERENCING STYLES - ELEMENTS

- Name of author
- Title of article/research paper
- Name of journal /periodical
- Volume number
- Date/month of issuance
- Year of publication
- Page number
- Name of book (in case of a book)
- Book publisher and place (in case of a book)

Referencing Styles APA (American Psychological Association) Referencing Style

The style to reference is 1. For a Research paper/Article Author's surname followed by his (their) initials (Year of publication) Article title. Name of Journal, Volume, Page no.

For Example: Tomaszewski, S. & Showerman, S. (2010). IFRS in United States: Challenges and opportunities, Review of Business,30,59-71

Referencing Styles APA (American Psychological Association)

Referencing Style, the style to reference is For a Book is

Author's surname, Initials (Year of publication) Book title. Place: Publisher Example Kumar, R (2014).

Research Methodology: A Step-by-Step Guide for Beginners (4th ed). Thousand Oaks, California: Sage Publications Referencing Styles Harvard Referencing Style The broad style to reference is the Author's Surname and initials. (Publication Year) 'Article title, Newspaper/Magazine Name, Day Month Published, Page(s). Available at: URL or DOI (Accessed date)

10.9 EVALUATION OF THE RESEARCH REPORT

A research report should be critically evaluated to obtain new insights and knowledge. The researcher could evaluate the report or get it done by somebody else. In certain situations, feedback comes automatically whereas in others the researcher may have to ask for it. Evaluation should be done objectively and preferably by people not associated with the research activities. A report can be evaluated against the following parameters:

(a) Does it address the problem correctly: A research report should be checked to see if it provides the right background to the problem and if it has successfully identified and located the problem. A report that does not provide such information is a weak report

(b) Is the research design appropriate: This question actually evaluates the report on two criteria; firstly, on a choice of research design and secondly on the description of the research design. The report should describe the method of drawing the sample collecting data and analysing it. However, the description should have been simple and non-technical. If the audience is unable to understand the research design, then the report falls a notch on the scale of a good report.

(c) Have appropriate numbers and statistics been used: Almost all reports make use of numbers and statistics to support discussion. These should be carefully examined to see if the appropriate statistics have been used and if they serve the purpose of the study well.

(d) Are the interpretations and conclusion objective: The report is evaluated for the objectivity and candidness of the interpretation of results. Any assumptions used while interpreting the results should have been clearly stated and the conclusions should be evaluated in the light of limitations faced by the researcher in his research process. This demands a complete and honest disclosure by the researcher.

(e) Are the results generalizable: The research report should be evaluated in terms of the generalizability of results. A good report provides sufficient evidence regarding the reliability, validity and generalizability of the findings. The target population to which the findings apply should be clearly identified and factors that limit the generalizability of results should be stated.

Thus, a good report would satisfy these questions that crop up in the mind of the reader to a great extent. An evaluator might develop a few other criteria of his own, but these questions are a good starting point to start evaluating the report.

10.10 QUESTIONS FOR PRACTICE

A. Short Answer Type Questions

Q1. What does interpretation mean?

Q2. Meaning of report writing in statistics.

Q3. Explain the meaning of a written report.

Q4. What is an oral report?

Q5. Give two precautions of interpretation.

B. Long Answer Type Questions

Q1. Explain benefits/ importance of interpretation

Q2. Discuss the precautions in interpretation

Q3. What are the qualities of a good research report

Q4. Explain the significance of report writing

Q5. What are the different types of reports?

Q6. What are the steps in writing the report

Q7. Explain the referencing styles.

Q8. Explain the steps to evaluate the research report

10.11 SUGGESTED READINGS

- Goode, W.J. and Hatt, P.K. (1952). Methods of Social Research. New York: McGraw Hill.
- Kothari, L.R. (1985). Research Methodology, New Delhi: Vishwa Prakashan.
- Anastas, J.W. (1999). Research Design for Social Work and The Human Services (2nd ed.) New York: Columbia University Press
- Burns, R.B. (2000). Introduction to Research Methods. New Delhi: Sage Publications.
- Ruane, J.M. (2005). Essentials of Research Methods: A guide to Social Science Research. Melbourne: Blackwell Publishing.