

E – NOTES

MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY



CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatISTICS & DATA SCIENCE

SUBJECT CODE: 23PEBC23

Programme Outcomes:	<p>PO1. To make students understand the importance of biochemistry as a subject that deals with life processes, as well as the concepts, theories and experimental approaches followed in biochemistry, in order to pursue a research career, either in an industry or academic setting.</p> <p>PO2. To develop analytical and problem-solving skills</p> <p>PO3. To create an awareness among the students on the interconnection between the interdisciplinary areas of biochemistry.</p> <p>PO4. To give the necessary practical skills required for biochemical techniques and analysis.</p> <p>PO5. To develop a communication and writing skills in students.</p> <p>PO6. To develop leadership and teamwork skills</p> <p>PO7. To emphasize the importance of good academic and work ethics and their social implications.</p> <p>PO8. To emphasize the importance of continuous learning and to promote lifelong learning and career development.</p> <p>PO9. To teach students how to retrieve information from a variety of sources, including libraries, databases and the internet.</p> <p>PO10. To teach students to identify, design and execute a research problem, analyze and interpret data and learn time and resource management.</p>
----------------------------	---

Course	CORE ELECTIVE PAPER –III
Title of the Course:	BIOSTATISTICS & DATA SCIENCE
Credits:	3
Pre-requisites, if any:	Basic knowledge of Statistics and Computer Applications
Course Objectives	<ol style="list-style-type: none"> 1. To summarize the data and to obtain its salient features from the vast mass of original data. 2. To understand the concept of various measures of dispersion. 3. To understand the concepts of sampling and learning test of significance. 4. To understand the concept of various attributes and relate to biological studies. 5. To gain knowledge in SPSS, a software package which gives a perfect graphical representation and appropriate result for the data that has been entered
Course Outcomes	<p>After completion of the course, the students should be able to:</p> <p>CO1: Concepts of statistical population and sample, variables and attributes. Tabular and graphical representation of data based on variables.(K1,K2,K3)</p> <p>CO2: Conditions for the consistency' and criteria for the independence of data based on attributes. Measures of central tendency, Dispersion, Skewness and Kurtosis.(K1,K2,K3)</p> <p>CO3: Learning different sampling methods and analysing statistical significance.(K1,K2,K3,K4)</p> <p>CO4: Understanding students t test , ANOVA , Chi square test to analyse the significance of various research. (K1,K2,K3,K4)</p> <p>CO5: Learning on data science, algorithm for machine learning, artificial intelligence and big data, their applications in clinical and pharma domain . (K1,K2,K3,K4.K6)</p>
Units	
I	Nature of biological and clinical experiments – Collection of data in experiment- Primary and secondary data. Methods of data collection. Classification and tabulation. Different forms of diagrams and graphs related to biological studies. Measures of Averages- Mean, Median, and mode. Use of these measures in biological studies.

II	Measures of Dispersion for biological characters – Quartile deviation, Mean deviation, Standard deviation and coefficient of variation. Measures of skewness and kurtosis. Correlation and regression – Rank correlation – Regression equation. Simple problems based on biochemical data.
III	Basic concepts of sampling- Simple random sample stratified sample and systemic sampling. Sampling distribution and standard error. Test of significance based on large samples. Test for mean, difference of means, proportions and equality of proportions.
IV	Small sample tests – Students‘t’ test for mean, difference of two way means, tests for correlation and regression coefficients. Chi-square test for goodness of a non independence of attributes. F test for equality of variances. ANOVA- one way and two way. Basic concept related to biological studies
V	Introduction to Data Science, Definition of data science, importance, and basic applications, Machine Learning Algorithms, Deep Learning, Artificial Neural Networks and their Application, Reinforcement Learning, Natural Language Processing Artificial Intelligence (AI), Data Visualization, Data Analysis, Optimization Techniques, Big Data, Predictive Analysis. Application of AI in medical, health and pharma industries.
Reading List (Print and Online)	<ol style="list-style-type: none"> 1. https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/Accessibility.pdf 2. https://pure.tue.nl/ws/portalfiles/portal/19478370/20160419_CO_Mzol_o.pdf 3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5453888/ 4. https://home.ubalt.edu/ntsbarsh/excel/excel.htm 5. https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_sps.pdf 6. https://www.ibm.com/support/pages/ibm-spss-statistics-28-documentation
Self-Study	<ol style="list-style-type: none"> 1. Simple problems on probability, theoretical distributions, hypothesis testing 2. Relationship between mean, median and mode pros and cons of the measures of central tendency and deviation
Recommended Texts	<ol style="list-style-type: none"> 1. Zar, J.H. (1984) “Bio Statistical Methods”, Prentice Hall, International Edition 2. Sundar Rao P. S.S., Jesudian G. & Richard J. (1987), “An Introduction to Biostatistics”, 2nd edition, Prestographik, Vellore, India,. 3. Warren, J; Gregory, E; Grant, R (2004), “Statistical Methods in Bioinformatics”, 1st edition, Springer 4. Milton, J.S. (1992), “Statistical methods in the Biological and Health Sciences”, 2nd edition, Mc Graw Hill, 5. Rosner, B (2005), “Fundamentals of Biostatistics”, Duxbury Press 6. Introducing Data Science, Davy Cielan, Anro DB Meysman, Mohamed Ali.

Method of Evaluation:

Test I	Test II	Assignment	End Semester Examination	Total	Grade
10	10	5	75	100	

Methods of assessment:

Recall (K1) - Simple definitions, MCQ, Recall steps, Concept definitions.

Understand/ Comprehend (K2) - MCQ, True/False, Short essays, Concept explanations, Short summary or overview.

Application (K3) - Suggest idea/concept with examples, Solve problems, Observe, Explain

Analyse (K4) - Problem-solving questions, Finish a procedure in many steps, Differentiate between various ideas

Evaluate (K5) - Longer essay/ Evaluation essay, Critique or justify with pros and cons

Create (K6) - Check knowledge in specific or off beat situations, Discussion, Presentations

Mapping with Programme Outcomes:

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10
CO 1	S	S	S	S	M	S	L	S	S	S
CO 2	S	S	S	S	M	S	L	S	S	S
CO 3	S	S	S	S	S	S	M	S	S	S
CO 4	S	S	S	S	S	S	M	S	S	S
CO 5	S	S	S	S	S	S	M	S	S	S

S-Strong

M-Medium

L-Low

Strong: 43

Medium: 05

Low:02

MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY

CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatISTICS & DATA SCIENCE

SUBJECT CODE: 23PEBC23

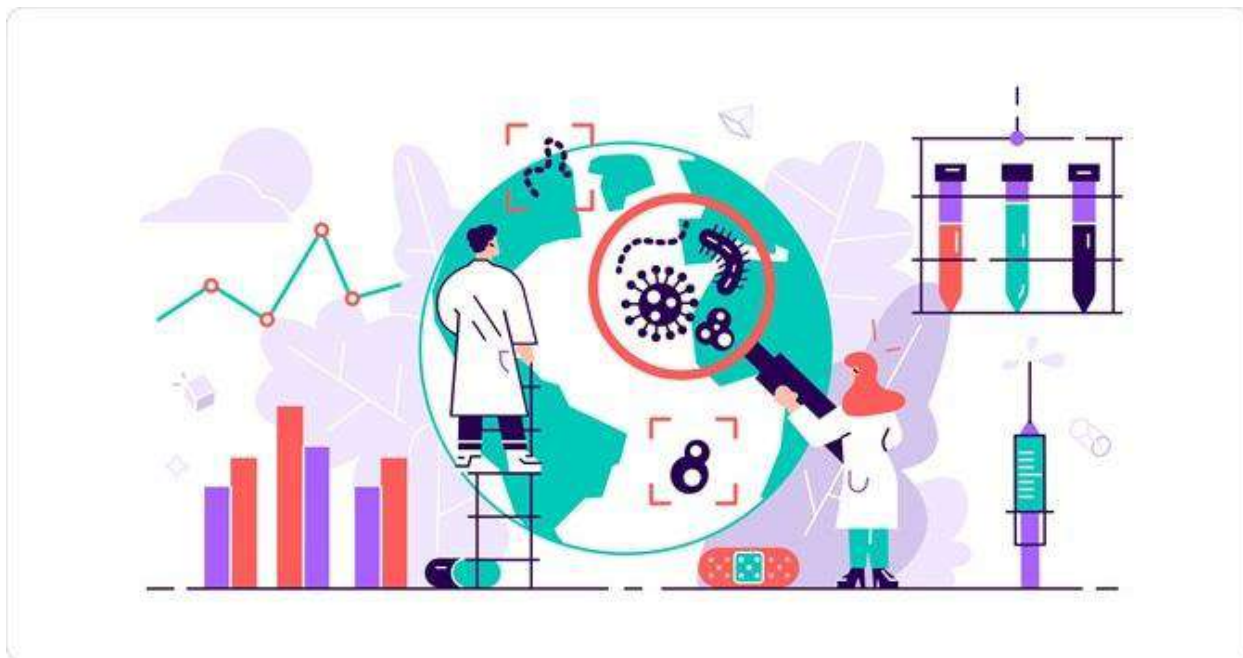
SYLLABUS

UNIT: I

UNIT I

Nature of biological and clinical experiments – Collection of data in experiment- Primary and secondary data. Methods of data collection. Classification and tabulation. Different forms of diagrams and graphs related to biological studies. Measures of Averages- Mean, Median, and mode. Use of these measures in biological studies.

BIostatISTICS & DATA SCIENCE



Dr. V. MAGENDIRA MANI, M.Sc., M.Phil., Ph.D., SET

Research Coordinator

PG and Research Department of Biochemistry

Marudhar Kesari Jain College for Women

Vaniyambadi-635 751

magendiramani@mkjc.in

STATISTICS

Statistics is a branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of data. It provides methods for making inferences about the characteristics of a population based on a limited set of observations, which is often referred to as a sample. Statistics is a fundamental tool in various fields, including science, business, economics, social sciences, and more.

Here are key concepts and components of statistics,

Descriptive Statistics

Measures of Central Tendency: These include the mean (average), median (middle value), and mode (most frequently occurring value). They provide a summary of the central or typical value of a dataset.

Measures of Dispersion: These include the range, variance, and standard deviation, which quantify the spread or variability of data points in a dataset.

Inferential Statistics

Hypothesis Testing: Statistical hypothesis testing is used to make inferences about a population based on a sample of data. It involves formulating a hypothesis, collecting data, and assessing whether the data provide enough evidence to reject or fail to reject the hypothesis.

Confidence Intervals: These provide a range of values within which a population parameter is likely to lie with a certain level of confidence.

Regression Analysis: This involves modeling the relationship between variables, allowing for prediction and understanding of the strength and nature of associations.

Probability

Probability Distributions: These describe the likelihood of different outcomes in a given set of events. Common probability distributions include the normal distribution, binomial distribution, and Poisson distribution.

Experimental Design

Randomization: Random assignment of subjects or treatments helps control for confounding variables in experimental studies.

Control Groups: Experimental design often involves the use of control groups to compare treatment effects against a baseline.

Statistical Software

****Tools like R, Python (with libraries like NumPy, SciPy, and Pandas), SAS, SPSS, and others are used for statistical analysis. These tools facilitate data manipulation, visualization, and complex statistical computations.**

Applications

Science: In scientific research, statistics help draw conclusions from experiments, assess the reliability of results, and make predictions.

Economics: In economic research, statistics are used to analyze trends, forecast future values, and evaluate the impact of policies.

Business: Businesses use statistics for market research, quality control, and decision-making processes.

Statistics is a versatile field, and its principles are applied in a wide range of disciplines to gain insights, make informed decisions, and draw meaningful conclusions from data.

BIOSTATISTICS

Biostatistics is a branch of statistics that involves the application of statistical methods to the field of biology and related disciplines. It encompasses the design, analysis, and interpretation of experiments and surveys that aim to collect, summarize, and draw inferences from biological data. Biostatistics plays a crucial role in various areas of biological and health sciences, contributing to the understanding of complex biological phenomena, the evaluation of medical treatments, and the formulation of public health policies.

Key components and applications of biostatistics include,

Study Design: Biostatisticians contribute to the planning and design of experiments and observational studies. They help researchers determine the appropriate sample size, randomization procedures, and data collection methods to ensure reliable and valid results.

Data Analysis: Biostatistical methods are used to analyze the collected data, ranging from descriptive statistics (e.g., means, medians) to inferential statistics (e.g., hypothesis testing, confidence intervals, regression analysis). This analysis aids in drawing conclusions about the relationships and patterns within the data.

Clinical Trials: Biostatistics plays a crucial role in the design and analysis of clinical trials, which are essential for evaluating the safety and efficacy of new medical treatments and interventions. Randomized controlled trials, a common study design in clinical research, rely on biostatistical principles to ensure the validity of results.

Epidemiology: Biostatistics is integral to epidemiological studies that investigate the distribution and determinants of health-related events in populations. These studies often involve the calculation of measures such as incidence, prevalence, and relative risk.

Public Health Research: Biostatistical methods are employed in public health research to assess the impact of various factors on health outcomes, study disease patterns, and inform public health interventions and policies.

Genetic Studies: In genetics and genomics research, biostatistics is used to analyze data from genome-wide association studies (GWAS), linkage analysis, and other genetic experiments. This aids in identifying genetic factors associated with diseases and traits.

Overall, biostatistics serves as a crucial tool in the scientific investigation of biological and health-related phenomena, providing quantitative methods to make sense of complex data and draw meaningful conclusions.

NATURE OF BIOLOGICAL AND CLINICAL EXPERIMENTS

Biological and clinical experiments are essential components of scientific research in the fields of biology and medicine. These experiments are designed to investigate various aspects of living organisms, understand biological processes, and improve our knowledge of health and disease. Here, I'll provide an overview of the nature of biological and clinical experiments:

Biological Experiments

Cellular and Molecular Biology

Aim: Understanding the fundamental processes within cells and at the molecular level.

Techniques: DNA sequencing, PCR (Polymerase Chain Reaction), Western blotting, ELISA (Enzyme-Linked Immunosorbent Assay), cell culture, microscopy.

Genetics

Aim: Studying inheritance patterns, gene function, and genetic variations.

Techniques: Genetic mapping, gene knockout studies, gene expression analysis, CRISPR-Cas9 technology.

Biochemistry

Aim: Investigating chemical processes within living organisms.

Techniques: Protein purification, enzyme assays, spectroscopy, chromatography.

Ecology

Aim: Understanding interactions between organisms and their environments.

Techniques: Field studies, population sampling, ecological modeling.

Neuroscience

Aim: Studying the structure and function of the nervous system.

Techniques: Brain imaging (MRI, fMRI), electrophysiology, behavioral experiments.

CLINICAL EXPERIMENTS

Clinical Trials

Aim: Assessing the safety and efficacy of new treatments or interventions in humans.

Phases: Phase I (safety), Phase II (efficacy), Phase III (large-scale efficacy), Phase IV (post-market surveillance).

Epidemiology

Aim: Investigating the distribution and determinants of health-related events in populations.

Designs: Cross-sectional studies, case-control studies, cohort studies.

Medical Imaging

Aim: Visualizing internal structures for diagnostic purposes.

Techniques: X-rays, CT scans, MRI, ultrasound, PET scans.

Observational Studies

Aim: Understanding natural history, patterns, and risk factors of diseases.

Types: Prospective (forward in time), retrospective (backward in time).

Interventional Studies

Aim: Evaluating the impact of specific interventions on health outcomes.

Examples: Randomized controlled trials (RCTs), clinical intervention studies.

Diagnostics

Aim: Developing and validating diagnostic tools and tests.

Techniques: PCR for infectious diseases, blood tests, imaging for early detection.

COMMON ELEMENTS**Hypothesis Testing**

Both types of experiments involve the formulation of hypotheses and testing them through systematic methods.

Data Collection and Analysis

Rigorous data collection and analysis are essential for drawing valid conclusions in both biological and clinical experiments.

Ethical Considerations

Adherence to ethical standards is crucial, especially in clinical experiments involving human subjects.

Peer Review

Results are often subjected to peer review to ensure the reliability and validity of findings before publication.

In summary, both biological and clinical experiments contribute to advancing our understanding of living organisms and improving healthcare outcomes. They often involve a combination of laboratory-based work, data analysis, and, in the case of clinical experiments, interactions with human subjects.

COLLECTION OF DATA

The collection of data in an experiment is a crucial step that involves gathering information or observations to answer specific research questions or test hypotheses. The process of data collection varies depending on the nature of the experiment, the research design, and the type of data required. Here are common methods and considerations for data collection in experiments:

PRIMARY DATA AND SECONDARY DATA

Primary data and secondary data are two types of data used in research and analysis, and they differ in their sources, collection methods, and purposes.

PRIMARY DATA

Primary data refers to information that is collected firsthand by the researcher specifically for the research project at hand. It is original and directly obtained from the source.

Sources

Surveys and Questionnaires: Researchers design and administer surveys to collect responses directly from individuals.

Interviews: One-on-one or group discussions where researchers gather information directly from participants.

Experiments: Conducting experiments to observe and collect data on variables of interest.

Observations: Directly observing and recording information about phenomena or behavior.

Advantages

Relevance: Data is tailored to the specific research objectives.

Accuracy: Researchers have control over the collection process.

Timeliness: Data is current and specific to the study.

Disadvantages

Cost and Time: Collecting primary data can be time-consuming and expensive.

Limited Scale: It may be challenging to collect data from a large population.

SECONDARY DATA

Definition

Secondary data refers to information that has already been collected by someone else for a different purpose but is used by the researcher for their own analysis.

Sources

Published Sources: Books, articles, reports, and other publications.

Government Reports: Data collected and published by government agencies.

Databases: Online databases and repositories containing pre-existing data.

Surveys and Studies: Data collected by other researchers for their projects.

Advantages

Cost and Time: Accessing existing data is often more cost-effective and quicker.

Large Scale: Secondary data may cover a larger population or time span.

Historical Analysis: Enables the study of trends and changes over time.

Disadvantages

Relevance: May not precisely meet the researcher's needs.

Quality Concerns: Data quality and accuracy are dependent on the source.

Unavailable Information: Some specific data may not be available.

When to Use Each**Primary Data:**

When specific information is required for the research objectives.

When existing data sources do not address the research questions.

In exploratory research or when the topic is novel.

Secondary Data

When cost and time constraints are significant considerations.

When historical trends or large-scale patterns are of interest.

In situations where primary data collection is not feasible.

In many research projects, a combination of both primary and secondary data may be used to leverage the advantages of each type and address the limitations. Researchers need to carefully consider their research goals, available resources, and the quality of data required when deciding between primary and secondary data sources.

METHODS OF DATA COLLECTION

Data collection is a crucial step in the research process, and researchers employ various methods to gather information for analysis. The choice of data collection method depends on the research question, the nature of the study, and the available resources. Here are some common methods of data collection:

Surveys and Questionnaires

Description: Researchers design a set of questions to gather information from participants.

Advantages: Cost-effective, can reach a large audience.

Considerations: Ensure clarity in questions, address potential bias, and maximize response rates.

Interviews

Description: Researchers ask questions directly to participants, either in person, over the phone, or through video calls.

Advantages: Allows for in-depth exploration, clarification of responses.

Considerations: Requires skilled interviewers, potential for interviewer bias.

Observations

Description: Systematic recording and analysis of behaviors, events, or phenomena.

Advantages: Provides direct, real-time information.

Considerations: Observer bias, ethical considerations, and potential reactivity.

Experiments

Description: Manipulating variables and measuring their effects to establish cause-and-effect relationships.

Advantages: Allows for control over variables, establishes causation.

Considerations: May lack external validity, ethical considerations.

Field Trials

Description: Testing interventions or treatments in real-world settings.

Advantages: Mimics real-world conditions.

Considerations: Less control than laboratory experiments, potential for confounding variables.

Case Studies

Description: In-depth analysis of a single individual, group, or phenomenon.

Advantages: Provides rich, detailed information.

Considerations: Limited generalizability, potential for bias.

Content Analysis

Description: Systematic analysis of textual, visual, or audio content.

Advantages: Allows for the study of media, documents, or communication.

Considerations: Requires clearly defined coding criteria.

Biometric Data Collection

Description: Measurement of physiological or biological parameters.

Examples: Heart rate monitoring, brain imaging, genetic testing.

Advantages: Objective measurements.

Considerations: Ethical concerns, potential for misinterpretation.

Sampling

Description: Selecting a subset of the population for study.

Examples: Random sampling, stratified sampling.

Advantages: Cost-effective, reduces the need to collect data from an entire population.

Considerations: Requires careful consideration of sampling methods.

Remote Sensing

Description: Collecting data from a distance using technology like satellites or sensors.

Examples: Satellite imagery, environmental sensors.

Advantages: Allows for large-scale data collection.

Considerations: Dependence on technology, potential for measurement errors.

Diaries and Logs

Description: Participants record their activities or experiences over time.

Advantages: Captures real-time data.

Considerations: Relies on participants' commitment and accuracy.

Internet and Social Media Analysis

Description: Analyzing data from online sources, social media platforms, or websites.

Advantages: Access to large amounts of user-generated data.

Considerations: Ethical considerations, potential for biased samples.

Researchers often use a combination of these methods to triangulate data and enhance the validity and reliability of their findings. The choice of method depends on the research objectives, the nature of the study, and practical considerations.

CONSIDERATIONS FOR EFFECTIVE DATA COLLECTION:**Validity**

Ensure that the data collected accurately measures what it is intended to measure.

Reliability

Aim for consistency and repeatability in data collection methods.

Ethical Considerations

Respect participant confidentiality, privacy, and informed consent.

Standardization

Use standardized protocols and procedures to minimize bias and enhance replicability.

Pilot Testing

Conduct a small-scale trial to identify and address potential issues before full-scale data collection.

Training

Train data collectors to ensure they follow standardized procedures and minimize errors.

Record Keeping

Maintain accurate and organized records of data to facilitate analysis and interpretation.

Data Security

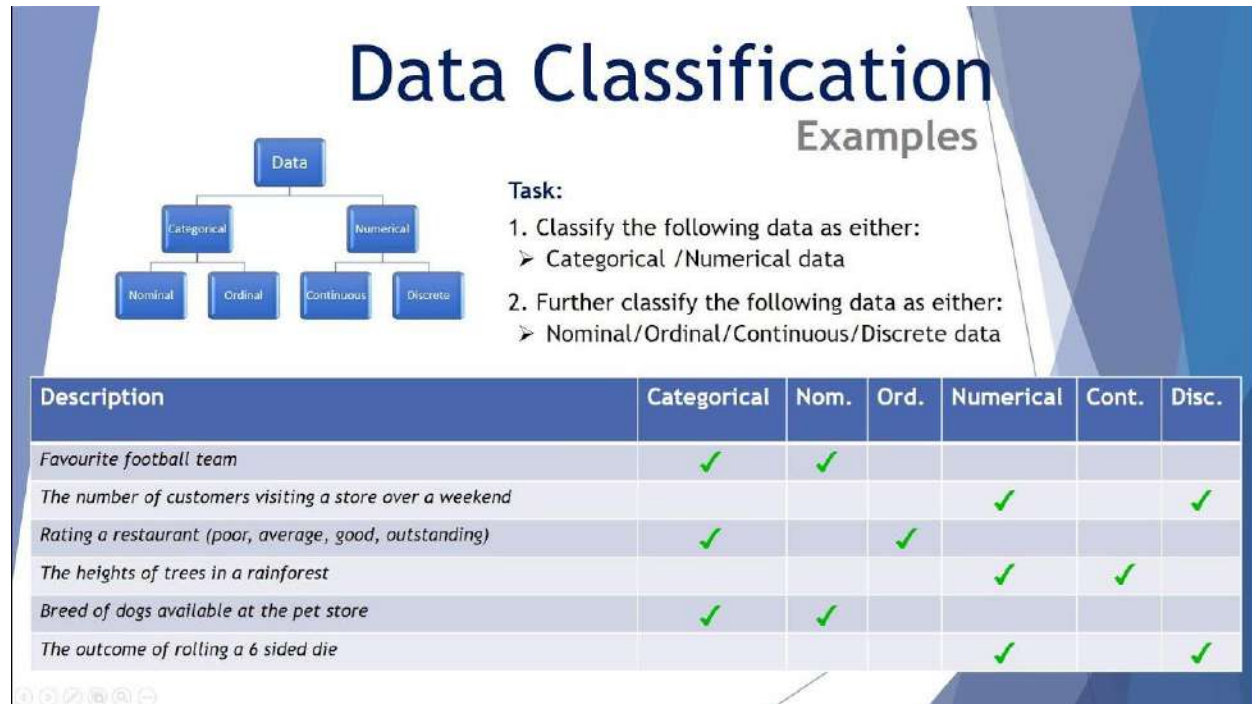
Implement measures to protect the integrity and confidentiality of collected data.

The effectiveness of an experiment often depends on the quality of the data collected. Careful planning, attention to detail, and adherence to ethical standards contribute to the success of the data collection process.

CLASSIFICATION AND TABULATION

Classification

Definition: Classification is the process of grouping and categorizing data or objects based on their similarities or common characteristics. It involves arranging diverse items into classes or categories to simplify and organize information.



The slide titled "Data Classification Examples" features a hierarchical diagram of data types, a task list, and a classification table.

Data Classification Hierarchy:

- Data
 - Categorical
 - Nominal
 - Ordinal
 - Numerical
 - Continuous
 - Discrete

Task:

1. Classify the following data as either:
 - Categorical / Numerical data
2. Further classify the following data as either:
 - Nominal/Ordinal/Continuous/Discrete data

Classification Table:

Description	Categorical	Nom.	Ord.	Numerical	Cont.	Disc.
<i>Favourite football team</i>	✓	✓				
<i>The number of customers visiting a store over a weekend</i>				✓		✓
<i>Rating a restaurant (poor, average, good, outstanding)</i>	✓		✓			
<i>The heights of trees in a rainforest</i>				✓	✓	
<i>Breed of dogs available at the pet store</i>	✓	✓				
<i>The outcome of rolling a 6 sided die</i>				✓		✓

Steps in Classification

- Identification of Characteristics: Identify the relevant characteristics or criteria based on which the classification will be done.
- Division into Groups: Group items or data points based on shared characteristics or attributes.
- Naming of Classes: Assign names or labels to the created groups to represent the categories.
- Consistency: Ensure consistency in applying criteria across different groups.

Example of Classification

Consider a classification of animals based on their habitats:

Class 1: Forest Animals (e.g., deer, bear, squirrel)

Class 2: Aquatic Animals (e.g., fish, dolphin, turtle)

Class 3: Desert Animals (e.g., camel, lizard, scorpion)

Tabulation

Definition: Tabulation involves the systematic arrangement of data in rows and columns, typically in a table. It provides a clear and organized presentation of information, making it easier to analyze and interpret.

Table Number

Title

Table 4.5

Population of India according to workers and non-workers by gender and location

Column Headings/Captions

(Crore)

Units

Location	Gender	Workers			Non-worker	Total
		Main	Marginal	Total		
Rural	Male	17	3	20	18	38
	Female	6	5	11	25	36
	Total	23	8	31	43	74
Urban	Male	7	1	8	7	15
	Female	1	0	1	12	13
	Total	8	1	9	19	28
All	Male	24	4	28	25	53
	Female	7	5	12	37	49
	Total	31	9	40	62	102

Source : Census of India 2001

Foot note : Figures are rounded to nearest crore

Source note

Footnote

Row Headings/stubs

Body of the table

Steps in Tabulation

- Determination of Columns and Rows: Decide on the variables or characteristics to be presented as columns and the individual items or categories as rows.
- Headings and Subheadings: Assign appropriate headings and subheadings to columns and rows for clarity.
- Numerical Entries: Fill in the table with numerical data corresponding to each intersection of rows and columns.
- Totals and Subtotals: Include totals and subtotals where necessary to summarize data.

Purpose

Classification: Organizes items based on similarities.

Tabulation: Presents data systematically for easy analysis.

Structure

Classification: Involves creating classes or groups.

Tabulation: Involves arranging data in a table format.

Representation

Classification: Often represented by categories or classes.

Tabulation: Represented by tables with rows and columns.

Use

Classification: Useful for organizing diverse data into manageable groups.

Tabulation: Useful for summarizing and presenting data in a structured format.

Both classification and tabulation are essential tools in statistics and data analysis, helping researchers and decision-makers make sense of complex information.

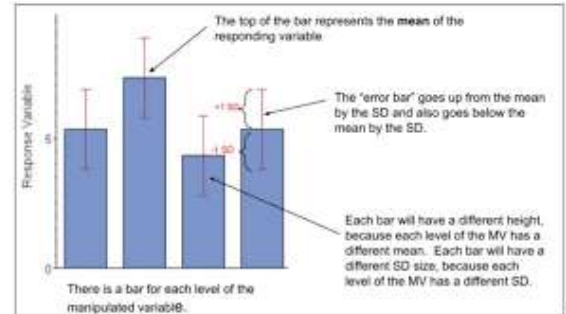
DIFFERENT FORMS OF DIAGRAMS AND GRAPHS RELATED TO BIOLOGICAL STUDIES.

Diagrams and graphs are widely used in biological studies to visually represent data, relationships, and concepts. They enhance the presentation of information and aid in the interpretation of complex biological phenomena. Here are some common forms of diagrams and graphs used in biological studies:

Bar Graphs

Purpose: Comparing categorical data or discrete groups.

Example: Comparing the number of individuals in different species in a habitat.



Histograms

Purpose: Representing the distribution of continuous data.

Example: Showing the frequency distribution of plant heights in a population.

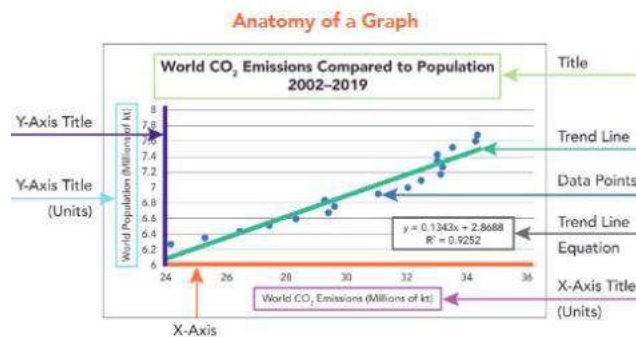


Line Graphs

Purpose: Displaying trends or changes in data over time or continuous variables.

Example: Illustrating changes in enzyme activity over a period of time.

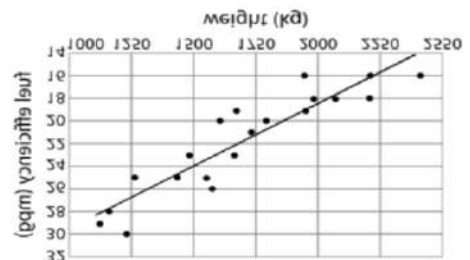
A graph is a way of visually representing the relationship between 2 or more variables.



Scatter Plots

Purpose: Showing the relationship between two continuous variables.

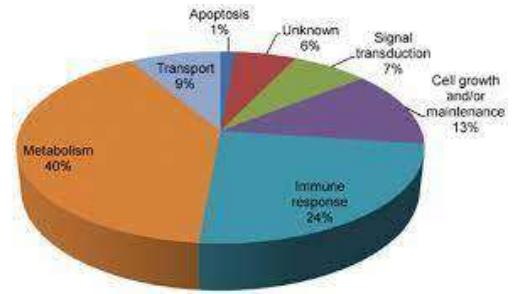
Example: Plotting the correlation between temperature and the growth rate of microorganisms.



Pie Charts

Purpose: Representing parts of a whole, displaying percentages.

Example: Showing the proportion of different macronutrients in a diet.



Venn Diagrams

Purpose: Illustrating the relationships between different sets.

Example: Comparing the characteristics of different animal species in terms of diet, habitat, and behavior.

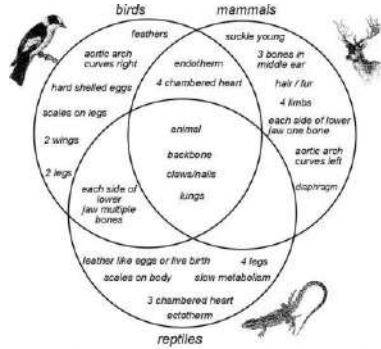
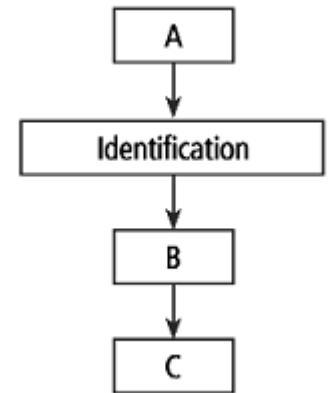


Figure 9.3 Venn diagrams of 3 classes of animals

Flowcharts

Purpose: Describing a process or sequence of events.

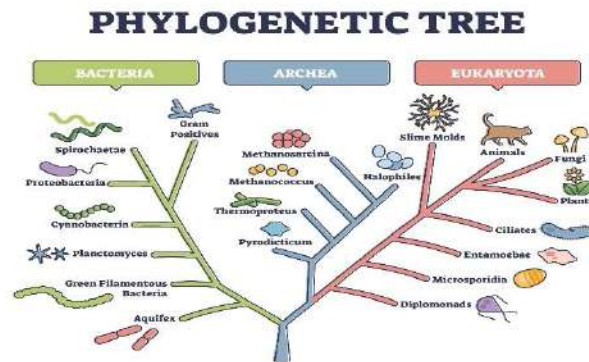
Example: Illustrating the steps involved in a biochemical pathway.



Phylogenetic Trees

Purpose: Showing the evolutionary relationships between species.

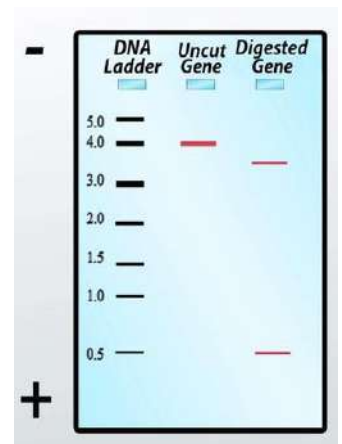
Example: Representing the evolutionary history of different organisms based on genetic data.



Electrophoresis Gel Images

Purpose: Displaying the separation of biomolecules (DNA, RNA, proteins) based on size or charge.

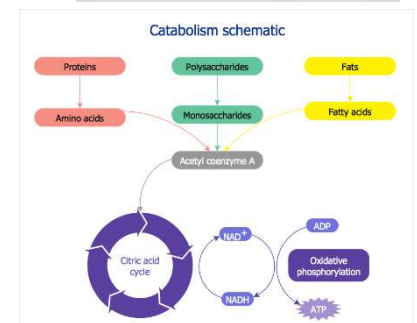
Example: Visualizing the results of DNA electrophoresis in a gel.



Schematic Diagrams

Purpose: Simplifying complex biological structures or processes.

Example: Illustrating the structure of a cell, including organelles and their functions.

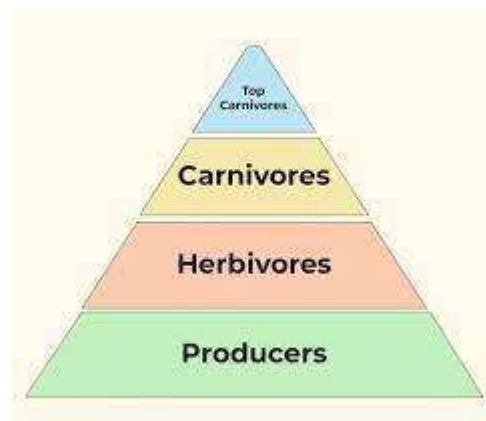
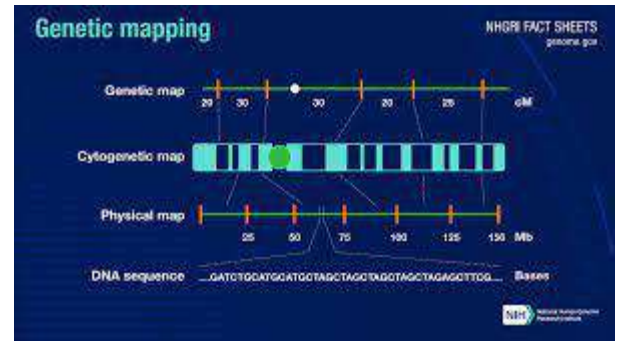


Purpose: Representing the arrangement of genes on a chromosome.

Example: Mapping the location of genes associated with a particular trait.

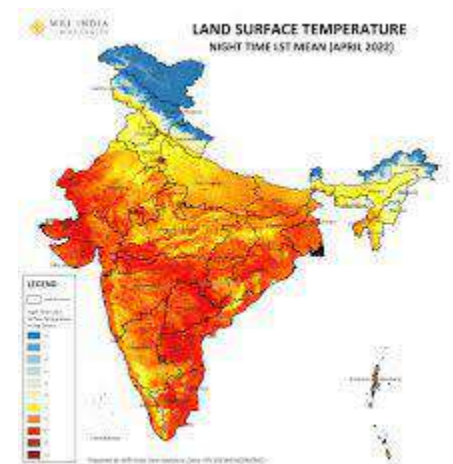
Purpose: Showing the trophic structure and energy flow in ecosystems.

Example: Representing the biomass or energy levels at each trophic level in a food chain.



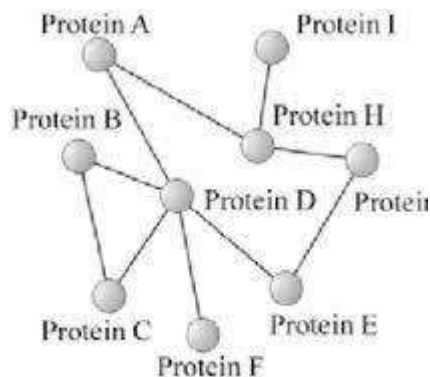
Purpose: Visualizing complex data matrices, such as gene expression levels.

Example: Displaying the expression patterns of genes across different experimental conditions.



Purpose: Illustrating interactions and relationships between biological entities.

Example: Representing protein-protein interactions or metabolic pathways.



(b)

When creating and interpreting diagrams and graphs in biological studies, it's crucial to choose the appropriate type based on the nature of the data and the message you want to convey. Clear and accurate visual representations enhance communication and facilitate a better understanding of biological concepts and findings.

MEASURES OF AVERAGES- MEAN, MEDIAN, AND MODE. USE OF THESE MEASURES IN BIOLOGICAL STUDIES

Measures of central tendency, such as mean, median, and mode, are widely used in biological studies to summarize and describe data. These measures provide insights into the typical or central value of a dataset, helping researchers understand the characteristics of a population. Here's how these measures are commonly used in biological studies:

Mean

Definition: The mean is the arithmetic average of a set of values calculated by adding all values and dividing by the number of observations.

Use in Biological Studies: The mean is frequently used to represent the average value of a quantitative variable. For example, in genetics, the mean might represent the average expression level of a gene across different samples. In ecology, the mean could represent the average size of a population.

Example

Find the **mean** of the set {2,5,5,6,8,8,9,11}.

Answer:

$$\text{mean} = \frac{2+5+5+6+8+8+9+11}{8} = 6.75$$

Median

Definition: The median is the middle value of a dataset when it is ordered. If there is an even number of observations, the median is the average of the two middle values.

Use in Biological Studies: The median is less sensitive to extreme values than the mean, making it useful when dealing with skewed distributions. In clinical studies, the median might be used to report the typical response time to a treatment, especially if response times are not normally distributed.

Find the **median** of the set {2,5,8,11,16,21,30}.

Answer:

There are seven numbers (ODD) in the set, and they are arranged in ascending order. The middle number (the 4th one in the list) is 11. so, the median is 11.

Find the **median** of the set {3,10,36,255,79,24,5,8}.

Answer:

Firstly, arrange the numbers in ascending order,

{3,5,8,10,24,36,79,255}

There are (8) numbers in the set (EVEN), so we have to find the two central numbers to calculate the median.

The middle numbers are {10,24}

So

$$\text{median} = \frac{10+24}{2} = 17$$

Mode

Definition: The mode is the value that occurs most frequently in a dataset.

Use in Biological Studies: The mode is used to identify the most common category or trait within a population. For example, in a study of plant morphology, the mode might represent the most frequently observed leaf shape. In medical research, the mode could indicate the most prevalent genotype in a population.

These measures are often used in combination to provide a more comprehensive description of the data. Additionally, measures of dispersion (e.g., standard deviation, range) are used alongside central tendency measures to give a fuller picture of the variability in biological datasets.

It's important for researchers to choose the appropriate measure based on the nature of the data and the research question. For instance, if the data is heavily skewed or contains outliers, the median may be a more appropriate measure of central tendency than the mean.

Example: The following table represents the number of wickets taken by a bowler in 10 matches. Find the mode of the given set of data.

Mode - mode of data

Match No.	1	2	3	4	5	6	7	8	9	10
No. of Wickets	2	1	1	3	2	3	2	2	4	1

It can be seen that 2 wickets were taken by the bowler frequently in different matches. Hence, the mode of the given data is 2.

MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY

CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatISTICS & DATA SCIENCE

SUBJECT CODE: 23PEBC23

SYLLABUS

UNIT: II

UNIT II

Measures of Dispersion for biological characters – Quartile deviation, Mean deviation, Standard deviation, and coefficient of variation. Measures of skewness and kurtosis. Correlation and regression – Rank correlation – Regression equation. Simple problems based on biochemical data.

MEASURES OF DISPERSION FOR BIOLOGICAL CHARACTERS

Measures of dispersion are used to describe the spread or variability of a set of data points. In the context of biological characters, these measures can help quantify how much individual observations vary from the central tendency (mean, median) of the data. Common measures of dispersion include:

1. Range:

- **Definition:** The difference between the maximum and minimum values in a dataset.
- **Formula:** $\text{Range} = \text{Max} - \text{Min}$

While the range is easy to compute, it is sensitive to outliers and may not provide a representative measure of dispersion in datasets with extreme values.

2. Variance:

- **Definition:** The average of the squared differences from the mean.
- **Formula for Population Variance:** $\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$
- **Formula for Sample Variance:** $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

Variance gives more weight to larger deviations from the mean and is useful for understanding the overall variability in a dataset.

3. Standard Deviation:

- **Definition:** The square root of the variance.
- **Formula for Population Standard Deviation:** $\sigma = \sqrt{\sigma^2}$
- **Formula for Sample Standard Deviation:** $s = \sqrt{s^2}$

Standard deviation is in the same units as the original data, making it more interpretable than variance.

4. Interquartile Range (IQR):

- **Definition:** The range of the middle 50% of the data, calculated as the difference between the third quartile (Q3) and the first quartile (Q1).
- **Formula:** $\text{IQR} = Q3 - Q1$

IQR is less sensitive to outliers than the range and provides a measure of the central spread in the middle of the distribution.

5. Coefficient of Variation (CV):

- **Definition:** The ratio of the standard deviation to the mean, expressed as a percentage.
- **Formula:** $CV = \left(\frac{s}{\bar{x}} \right) \times 100$

CV allows for the comparison of variability between datasets with different units or scales.

Choose the measure of dispersion that best suits your data and research objectives. Different measures have different strengths and weaknesses, and the choice may depend on the characteristics of your biological data.

EXAMPLE

Range

Definition:

The difference between the maximum and minimum values in a dataset.

Example:

If you are measuring the height of a population of plants, the range would be the difference between the tallest and shortest plant.

Variance

Definition:

The average of the squared differences from the mean.

Example:

If you have a dataset of animal weights (in kilograms), you would calculate the variance to understand how much each individual animal's weight deviates from the mean weight.

Standard Deviation

Definition:

The square root of the variance.

Example:

In a study of fish lengths (in centimeters), the standard deviation would provide a measure of how spread out the lengths are from the mean length.

Interquartile Range (IQR)

Definition:

The range of the middle 50% of the data, calculated as the difference between the third quartile (Q3) and the first quartile (Q1).

Example:

If you are examining the flowering time of a population of plants, the IQR would give you the range of time during which the middle 50% of the plants flower.

Coefficient of Variation (CV)

Definition: The ratio of the standard deviation to the mean, expressed as a percentage.

Example: If you are comparing the variability in the sizes of two populations of insects, the coefficient of variation would help you assess which population has more relative variability.

QUARTILE DEVIATION

Quartile Deviation (also known as Semi-Interquartile Range) is a measure of statistical dispersion that describes the spread of a dataset by indicating the range within which the middle half of the observations lie. It is based on quartiles, specifically the interquartile range (IQR).

Here are the steps to calculate the Quartile Deviation:

Calculate the First Quartile (Q1):

- Q1 is the median of the lower half of the data set.
- If the data set has an odd number of observations, Q1 is the middle value.
- If the data set has an even number of observations, Q1 is the average of the two middle values.

2. Calculate the Third Quartile (Q3):

- Q3 is the median of the upper half of the data set.
- If the data set has an odd number of observations, Q3 is the middle value.
- If the data set has an even number of observations, Q3 is the average of the two middle values.

3. Calculate the Interquartile Range (IQR):

- IQR is the difference between Q3 and Q1.
- $IQR = Q3 - Q1$

4. Calculate the Quartile Deviation:

- Quartile Deviation is half of the interquartile range.
- $Quartile\ Deviation = \frac{IQR}{2}$

The Quartile Deviation provides a measure of the spread of the central half of the dataset. It is less sensitive to extreme values compared to the range and standard deviation, making it a more robust measure of dispersion in the presence of outliers.

Formulaically, if QD is the quartile deviation and $Q3$ and $Q1$ are the third and first quartiles respectively:

$$QD = \frac{Q3 - Q1}{2}$$

This measure is particularly useful when dealing with skewed distributions or datasets with outliers, as it focuses on the variability in the middle of the distribution rather than being heavily influenced by extreme values.

Suppose we have a dataset representing the scores of 10 students in a biology exam:

60, 65, 70, 72, 75, 78, 80, 82, 85, 90

1. Calculate Quartiles:

- First Quartile (Q1): $60 + (0.25 \times (65 - 60)) = 61.25$
- Third Quartile (Q3): $75 + (0.75 \times (78 - 75)) = 76.25$

2. Calculate Quartile Deviation:

- Quartile Deviation (QD): $\frac{76.25 - 61.25}{2} = 7.5$

So, the Quartile Deviation for this dataset is 7.5.

This means that the middle 50% of the data (between Q1 and Q3) has an average spread of 7.5 units. The Quartile Deviation provides a measure of variability that is less affected by extreme values than some other measures of dispersion, such as the standard deviation.

EXAMPLE - MEAN DEVIATION

Suppose we have a dataset representing the daily temperatures (in degrees Celsius) over a week:

20, 22, 18, 25, 21, 23, 19

1. Calculate the Mean:

- $\bar{X} = \frac{20+22+18+25+21+23+19}{7} = \frac{148}{7} \approx 21.14$

2. Calculate Mean Deviation:

- Mean Deviation = $\frac{|20-21.14|+|22-21.14|+\dots+|19-21.14|}{7}$
- Mean Deviation = $\frac{1.14+0.86+\dots+2.14}{7}$
- Mean Deviation $\approx \frac{6.57}{7} \approx 0.94$

So, the Mean Deviation for this temperature dataset is approximately 0.94 degrees Celsius. It provides an average measure of how much each temperature reading deviates from the mean.

MEAN DEVIATION

The mean deviation (also known as the average absolute deviation) is a measure of the dispersion or spread of a set of values. It calculates the average absolute difference between each data point and the mean of the dataset.

Here are the steps to calculate the mean deviation:

1. Calculate the Mean (\bar{x}):

- Add up all the values in the dataset and divide by the number of values (n).
- $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

2. Calculate the Absolute Deviation for each Data Point:

- Find the absolute difference between each data point (x_i) and the mean (\bar{x}).
- $|x_i - \bar{x}|$

3. Calculate the Mean Deviation:

- Find the average of all the absolute deviations.
- Mean Deviation = $\frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$

In formulaic terms, the mean deviation (MD) is given by:

$$MD = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

The mean deviation is a simple measure of dispersion and is less affected by extreme values (outliers) than the standard deviation. However, it tends to underestimate the true variability in the data because it only considers the absolute differences without regard to the direction (above or below the mean).

While the mean deviation is a valid measure, the standard deviation is more commonly used in statistical analysis because it has certain mathematical properties that make it preferable in many contexts. Nonetheless, the mean deviation can be a useful alternative, especially when you want a measure of dispersion that is less sensitive to extreme values.

STANDARD DEVIATION AND COEFFICIENT OF VARIATION

The standard deviation and coefficient of variation are both measures of dispersion used in statistics to quantify the spread or variability of a dataset. They provide insights into how individual data points deviate from the central tendency (mean) of the dataset.

Standard Deviation:

Definition: The standard deviation is a measure of the amount of variation or dispersion in a set of values. It is calculated as the square root of the variance.

- **Formula for Population Standard Deviation:** $\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}}$
- **Formula for Sample Standard Deviation:** $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$

The standard deviation is expressed in the same units as the original data and provides a more interpretable measure of dispersion. It is commonly used and is sensitive to the magnitudes of deviations from the mean.

2. Coefficient of Variation (CV):

- **Definition:** The coefficient of variation is a relative measure of dispersion that expresses the standard deviation as a percentage of the mean. It is useful for comparing the relative variability of datasets with different units or scales.
- **Formula:** $CV = \left(\frac{s}{\bar{x}}\right) \times 100$

The coefficient of variation is particularly valuable when comparing the variability of datasets with different means. A lower CV indicates lower relative variability, while a higher CV suggests higher relative variability.

Comparison:

- The standard deviation gives an absolute measure of dispersion in the same units as the original data.
- The coefficient of variation provides a relative measure that is unitless and is useful for comparing the relative variability of datasets with different means.

Example:

- If Dataset A has a standard deviation of 10 and a mean of 50, the coefficient of variation is $\left(\frac{10}{50}\right) \times 100 = 20\%$.
- If Dataset B has a standard deviation of 5 and a mean of 25, the coefficient of variation is $\left(\frac{5}{25}\right) \times 100 = 20\%$.

In this example, both datasets have the same  coefficient of variation (20%), indicating similar relative variability despite having different means and standard deviations.

MEASURES OF SKEWNESS AND KURTOSIS

Skewness and kurtosis are two statistical measures that provide insights into the shape of a probability distribution.

1. Skewness:

Skewness measures the asymmetry of a probability distribution. It indicates whether the data is skewed to the left or right.

A positive skewness indicates a right-skewed distribution, where the right tail is longer or fatter than the left tail. In other words, the majority of the data points are concentrated on the left side.

A negative skewness indicates a left-skewed distribution, where the left tail is longer or fatter than the right tail. In this case, the majority of the data points are concentrated on the right side.

- Skewness is typically calculated using the formula:

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

where n is the number of observations, x_i is each individual observation, \bar{x} is the mean, and s is the standard deviation.

2. Kurtosis:

- Kurtosis measures the "tailedness" of a probability distribution. It provides information about the thickness of the tails relative to the normal distribution.
- A positive kurtosis (leptokurtic) indicates heavy tails, meaning that the distribution has more extreme values than a normal distribution. The peak of the distribution is higher and sharper.
- A negative kurtosis (platykurtic) indicates light tails, meaning that the distribution has fewer extreme values than a normal distribution. The peak of the distribution is lower and broader.
- Kurtosis is typically calculated using the formula:

$$\text{Kurtosis} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

In both formulas, n is the number of observations, x_i is each individual observation, \bar{x} is the mean, and s is the standard deviation.

It's important to note that there are different versions of skewness and kurtosis formulas, and the ones mentioned here are based on the moment method. There are also sample skewness and kurtosis formulas that correct for bias in small sample sizes.

CORRELATION AND REGRESSION

Correlation and regression are two statistical techniques used to analyze the relationship between two or more variables.

1. Correlation:

- Correlation measures the strength and direction of a linear relationship between two variables. The correlation coefficient, often denoted by r , ranges from -1 to 1.
- A positive correlation ($r > 0$) indicates a direct or positive relationship: as one variable increases, the other tends to increase as well.
- A negative correlation ($r < 0$) indicates an inverse or negative relationship: as one variable increases, the other tends to decrease.
- A correlation coefficient of 0 implies no linear relationship between the variables.
- The formula for the Pearson correlation coefficient (r) between variables X and Y is given by:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

where X_i and Y_i are individual data points, and \bar{X} and \bar{Y} are the means of variables X and Y .

2. Regression:

- Regression analysis is used to model the relationship between a dependent variable (Y) and one or more independent variables (X_1, X_2, \dots, X_n). It helps to predict the value of the dependent variable based on the values of the independent variables.
- In simple linear regression, with one independent variable X and one dependent variable

- In simple linear regression, with one independent variable X and one dependent variable Y , the relationship is represented by the equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where β_0 is the intercept, β_1 is the slope (regression coefficient), and ε is the error term.

- The coefficients (β_0 and β_1) are estimated using the least squares method to minimize the sum of squared differences between the observed and predicted values of Y .
- Multiple linear regression extends this concept to more than one independent variable.

In summary, correlation measures the strength and direction of a linear relationship between two variables, while regression models and quantifies the relationship, allowing predictions to be made based on the values of one or more independent variables.

RANK CORRELATION

Rank correlation is a statistical technique used to measure the degree of association between two variables by comparing their rankings. It is particularly useful when the data are not on an interval or ratio scale and may not meet the assumptions of parametric correlation methods like Pearson correlation.

The two commonly used rank correlation coefficients are Spearman's rank correlation coefficient (ρ) and Kendall's tau (τ).

1. Spearman's Rank Correlation Coefficient (ρ):

- Spearman's rank correlation is used when the variables are measured on an ordinal scale.
- It assesses the strength and direction of monotonic relationships between the ranks of two variables.
- The formula for Spearman's rank correlation is:

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

where d_i is the difference between the ranks of corresponding pairs of observations and n is the number of pairs.


2. Kendall's Tau (τ):

- Kendall's tau is also used for ordinal data and measures the degree of similarity in the ordering of data points between two variables.
- It is based on the count of concordant and discordant pairs of observations.
- The formula for Kendall's tau is:

$$\tau = \frac{\text{Number of concordant pairs} - \text{Number of discordant pairs}}{\frac{n(n-1)}{2}}$$

where n is the number of pairs.

Both Spearman's rank correlation coefficient and Kendall's tau range from -1 to 1. A positive value indicates a positive association (i.e., as one variable increases, the other tends to increase), while a negative value indicates a negative association. A value of 0 suggests no monotonic relationship.

These rank correlation coefficients are non-parametric methods, meaning they do not rely on the distributional assumptions of the data.  They are suitable for a wider range of variable types, including ordinal data.

REGRESSION EQUATION

A regression equation is a mathematical model that describes the relationship between a dependent variable and one or more independent variables. In simple linear regression, there is one independent variable, while in multiple linear regression, there are two or more independent variables. The goal of regression analysis is to find the best-fitting line (or hyperplane in the case of multiple variables) that minimizes the difference between the observed values and the values predicted by the model.

1. Simple Linear Regression:

”

The equation for a simple linear regression model can be expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Y is the dependent variable.
- X is the independent variable.
- β_0 is the y-intercept (the value of Y when X is 0).
- β_1 is the slope of the regression line (the change in Y for a one-unit change in X).
- ε represents the error term, accounting for unexplained variability.

2. Multiple Linear Regression:

In the case of multiple linear regression with k independent variables, the equation is extended as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

- Y is still the dependent variable.
- X_1, X_2, \dots, X_k are the independent variables.
- β_0 is the y-intercept.
- $\beta_1, \beta_2, \dots, \beta_k$ are the coefficients that represent the change in Y for a one-unit change in the corresponding X .
- ε represents the error term.

The goal of regression analysis is to estimate the values of the coefficients ($\beta_0, \beta_1, \dots, \beta_k$) that best fit the observed data. This is typically done by minimizing the sum of squared differences between the observed values and the values predicted by the regression equation. The resulting equation can then be used to make predictions or understand the relationship between variables.

EXAMPLE

Let's consider a simple example of simple linear regression. Suppose we have data on the number of hours students spend studying (X) and their exam scores (Y).

$$X = [2, 3, 4, 5, 6]$$

$$Y = [60, 65, 75, 80, 85]$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

We'll use statistical methods to estimate the coefficients β_0 and β_1 .

1. Calculate the means:

$$\bullet \bar{X} = \frac{2+3+4+5+6}{5} = 4$$

$$\bullet \bar{Y} = \frac{60+65+75+80+85}{5} = 73$$

2. Calculate the slope (β_1):

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\beta_1 = \frac{(2-4)(60-73) + (3-4)(65-73) + \dots + (6-4)(85-73)}{(2-4)^2 + (3-4)^2 + \dots + (6-4)^2}$$

After calculation, $\beta_1 \approx 5$.

3. Calculate the y-intercept (β_0):

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$$\beta_0 = 73 - (5 \times 4) = 53$$

So, the regression equation for this example is:

$$Y = 53 + 5X + \varepsilon$$

This equation can be used to predict exam scores based on the number of hours studied. For instance, if a student studies for 7 hours ($X = 7$), the predicted exam score would be $53 + 5 \times 7 = 88$.

MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY

CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatISTICS & DATA SCIENCE

SUBJECT CODE: 23PEBC23

SYLLABUS

UNIT: III

UNIT III

Basic concepts of sampling- Simple random sample stratified sample and systemic sampling. Sampling distribution and standard error. Test of significance based on large samples. Test for mean, difference of means, proportions and equality of proportions.

BASIC CONCEPTS OF SAMPLING

Sampling is a crucial technique in statistics that involves selecting a subset of individuals or elements from a larger population for the purpose of making inferences about the population based on the characteristics of the sample. Here are some basic concepts of sampling:

Population

The population refers to the entire group of individuals or elements that share a common characteristic and are the subject of the study. It's the larger group from which the sample is drawn.

Sample

A sample is a subset of the population selected for the study. The goal is for the sample to be representative of the larger population to make valid inferences.

Sampling Frame

The sampling frame is a list or representation of the elements in the population from which the sample will be drawn. It's important that the sampling frame accurately represents the population.

Sampling Methods

There are various methods for selecting a sample from a population, including:

SIMPLE RANDOM SAMPLING

A simple random sample is a type of probability sampling method where each member of the population has an equal chance of being selected, and the selection of one individual does not influence the selection of another. This is a straightforward and unbiased way of selecting a sample from a larger population.

Example

Let's consider a simple random sampling example involving a small population of students in a school. Suppose the population consists of 50 students, and we want to select a simple random sample of 10 students for a survey on study habits.

Steps:

Define the Population:

The population is all 50 students in the school.

Create a Sampling Frame:

A list of all 50 students' names serves as the sampling frame.

Random Selection:

Use a random method to select 10 students from the sampling frame. For simplicity, let's use random number generation. Each student is assigned a number from 1 to 50, and we randomly select 10 numbers.

Randomly selected numbers: 7, 12, 18, 25, 31, 36, 40, 45, 48, 50

Equal Probability:

Each student in the sampling frame has an equal probability of being selected.

Without Replacement:

Once a student is selected, they are not returned to the sampling frame for the subsequent selections.

Result:

The selected students based on the random numbers are Student 7, Student 12, Student 18, Student 25, Student 31, Student 36, Student 40, Student 45, Student 48, and Student 50.

These 10 students constitute a simple random sample from the population of 50 students. Researchers can now survey these students to gather information about study habits, and the findings can be generalized to the entire population, assuming that the sampling was done with care and the sampling frame accurately represents the population.

STRATIFIED SAMPLING

Stratified sampling is a method of sampling that involves dividing the population into subgroups or strata based on certain characteristics, and then taking a random sample from each stratum. This technique ensures that each subgroup is adequately represented in the final sample, allowing for more precise analysis of each stratum and often providing more accurate overall estimates for the entire population.

Example:

Consider a university with a total student population of 1,000 students. The students can be stratified based on their academic majors into three strata: Science, Arts, and Business.

Define the Population:

All 1,000 students in the university.

Identify Strata:

Three strata: Science majors, Arts majors, and Business majors.

Homogeneous Strata:

Students within each stratum (Science, Arts, Business) should have similar academic majors.

Random Sampling within Strata:

Randomly select, for example, 20% of students from each stratum. This ensures that each major is represented proportionally in the final sample.

Combine Samples:

Combine the samples from each major to form the overall stratified sample.

Stratified sampling is particularly useful when there is significant variability within the population, and the researcher wants to ensure representation from various subgroups. It can result in more precise and reliable estimates for each stratum and, consequently, for the entire population.

SYSTEMATIC SAMPLING

Elements are selected at regular intervals from a list after a random starting point. Systematic sampling is a method of sampling where every N^{th} item in the population is selected after starting from a randomly chosen initial item. It is a systematic and straightforward way to obtain a sample from a larger population, often with less effort than simple random sampling.

Example:

Consider a factory with 500 employees, and the goal is to select a systematic sample of 50 employees to conduct a job satisfaction survey.

Define the Population:

All 500 employees in the factory.

Determine the Sample Size:

Decide on a sample size, let's say 50 employees.

Calculate Sampling Interval (k):

$$k = \frac{500}{50} = 10$$

Randomly Select a Starting Point:

Choose a random number between 1 and 10. Let's say the random starting point is 7.

Select Every 10th Employee:

Include the 7th, 17th, 27th, ..., 497th, and 500th employees in the sample until you reach the desired sample size of 50.

Systematic sampling is relatively easy to implement and is more practical than simple random sampling in certain situations. However, it may introduce bias if there is a periodic pattern in the population that aligns with the sampling interval. For example, if there is a weekly cycle in a factory, and the sampling interval is a multiple of 7, systematic sampling might not be the best choice.

SAMPLING DISTRIBUTION


A sampling distribution is a statistic that determines the probability of an event based on data from a small group within a large population.

Sampling distributions are created by drawing many random samples of a given size from the same population. They help you understand how a sample statistic varies from sample to sample.

Sampling distributions describe a range of possible outcomes for a statistic, such as the mean or mode of some variable, of a population.

The central limit theorem (CLT) states that the distribution of a sample variable approximates a normal distribution (i.e., a “bell curve”) as the sample size becomes larger.

The standard deviation of the sampling distribution measures how much the sample statistic varies from sample to sample. It is smaller than the standard deviation of the population by a factor of \sqrt{n}



Sampling Distribution

['sam-plɪŋ di-strə-'byü-shən]

A probability distribution of a statistic obtained from a larger number of samples drawn from a specific population.

Practical Example

Suppose you want to find the average height of children at the age of 10 from each continent. You take random samples of 100 children from each continent, and you compute the mean for each sample group.

For example, in South America, you randomly select data about the heights of 10-year-old children, and you calculate the mean for 100 of the children. You also randomly select data from North America and calculate the mean height for one hundred 10-year-old children.

As you continue to find the average heights for each sample group of children from each continent, you can calculate the mean of the sampling distribution by finding the mean of all the average heights of each sample group. Not only can it be computed for the mean, but it can also be calculated for other statistics such as standard deviation and variance.

SAMPLING ERROR

Sampling error is the deviation between a sample (the mean or proportion) and the corresponding population parameter. Reducing it aims to improve statistical estimates' accuracy and reliability and minimize the risk of making incorrect inferences about the population.

Sampling Error



Example #1

Suppose a researcher wants to estimate the average height of adult males in a city. They randomly sample 100 males from a population of 100,000 and calculate the mean sample height as 5 feet 10 inches. The researcher then uses this sample mean to estimate the population means size, assuming that the model represents the people.

However, the error associated with this estimate is likely to be large, given the small sample size relative to the population size. Furthermore, if the standard deviation of heights in the population is high, the error could be quite large and lead to incorrect inferences about the population's mean size.

Example #2

One example of a recent sampling error in the news is the COVID-19 vaccine efficacy estimates. Vaccine efficacy estimates are calculated by comparing the number of COVID-19 cases in the vaccinated group to those in the placebo group. However, these estimates are subject to such error due to the small sample sizes and the random variation in the number of issues between the two groups.

For example, the efficacy estimate for the Pfizer vaccine was initially reported to be 95%, but this estimate had a confidence interval that ranged from 90% to 98%. This means that the true vaccine efficacy may be lower or higher than the reported estimate due to chance variation in the sample.

For the sample mean (\bar{X}):

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}$$

where:

- s is the sample standard deviation,
- n is the sample size.

For the sample proportion (\hat{p}):

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where:

\hat{p} is the sample proportion,
 n is the sample size.

Key points about the standard error:

Precision of the Estimate:

A smaller standard error indicates a more precise estimate of the population parameter. In other words, it suggests that the sample statistic is likely to be closer to the true population parameter.

Relationship with Sample Size:

As the sample size (n) increases, the standard error tends to decrease. Larger sample sizes lead to more reliable and precise estimates.

Use in Confidence Intervals:

Standard error is often used to calculate confidence intervals. The margin of error in a confidence interval is determined by the standard error.

Use in Hypothesis Testing:

In hypothesis testing, the standard error is used to calculate test statistics, such as the t-statistic or z-statistic, which are then used to determine the significance of an observed effect.

Different Formulas for Different Statistics:

The formula for standard error depends on the specific sample statistic being considered. The examples provided above are for the sample mean and sample proportion.

In summary, standard error is a crucial concept in inferential statistics, providing a measure of the variability or uncertainty associated with sample statistics. It is widely used in constructing confidence intervals, performing hypothesis tests, and assessing the reliability of sample estimates.

TEST OF SIGNIFICANCE BASED ON LARGE SAMPLES

When dealing with large samples, statistical tests of significance often rely on the Central Limit Theorem (CLT). The CLT states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables.

Commonly used tests for significance with large samples include:

Z-Test for a Population Mean:

Used when you know the population standard deviation (σ) and are testing the mean of a sample.

Formula:

$$\text{Formula: } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Example: Testing if the average height of a sample differs significantly from a known population mean.

Average Weight of a Population:

Scenario: A researcher wants to investigate whether the average weight of a certain population of mice is significantly different from a known average weight.

Test: Z-Test for a Population Mean.

Hypothesis:

Null Hypothesis (H_0): The average weight of the population is equal to the known average weight.

Alternative Hypothesis (H_1): The average weight of the population is not equal to the known average weight.

Analysis: Collect a large sample of mice, measure their weights, and conduct a Z-Test for a Population Mean.

TEST FOR MEAN

There are several statistical tests that can be used to assess whether the mean of a sample is significantly different from a hypothesized population mean. Here are a couple of common ones:

One-Sample t-Test:

Objective: To test whether the mean of a single sample is significantly different from a known or hypothesized population mean.

Assumptions: The data should be approximately normally distributed, and the observations should be independent.

Formula:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Where:

- \bar{x} is the sample mean.
- μ is the population mean.
- s is the sample standard deviation.
- n is the sample size.

Z-Test:

Objective: Similar to the one-sample t-test, but it's used when the population standard deviation is known.

Assumptions: The data should be approximately normally distributed, and the observations should be independent.

Formula:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Where:

- \bar{x} is the sample mean.
- μ is the population mean.
- σ is the population standard deviation.
- n is the sample size.

Steps for Hypothesis Testing:**Formulate Hypotheses:**

H0 : Null hypothesis (usually stating no effect or no difference).

H1 or H α : Alternative hypothesis (claiming an effect or difference).

Choose Significance Level (α):

Common choices are 0.05, 0.01, etc.

Collect and Analyze Data:

Collect your sample data.

Calculate the test statistic.

Make a Decision:

If the p-value is less than or equal to α , reject the null hypothesis.

If the p-value is greater than α , fail to reject the null hypothesis.

Draw a Conclusion:

Based on your decision, draw a conclusion in the context of the problem.

DIFFERENCE OF MEANS

When you're interested in comparing the means of two independent groups, there are statistical tests designed for this purpose. Here are two commonly used tests:

Independent Samples t-Test:

Objective: To compare the means of two independent groups to determine if they are significantly different from each other.

Assumptions: The data in each group should be approximately normally distributed, and the observations in each group should be independent.

Formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- \bar{x}_1 and \bar{x}_2 are the sample means of the two groups.
- s_1 and s_2 are the sample standard deviations of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

Paired Samples t-Test (for Dependent Samples):

Objective: To compare the means of two related groups (paired observations, e.g., before and after treatment) to determine if there's a significant difference.

Assumptions: The differences between paired observations should be approximately normally distributed.

Formula:

$$t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Where:

- \bar{d} is the mean of the differences between paired observations.
- s_d is the standard deviation of the differences.
- n is the number of paired observations.

Steps for Hypothesis Testing (t-Test):**Formulate Hypotheses:**

H_0 : Null hypothesis (usually stating no difference in means).

H_1 or H_a : Alternative hypothesis (claiming a significant difference in means).

Choose Significance Level (α):

Common choices are 0.05, 0.01, etc.

Collect and Analyze Data:

Collect your sample data.

Calculate the test statistic.

Make a Decision:

If the p-value is less than or equal to α , reject the null hypothesis.

If the p-value is greater than α , fail to reject the null hypothesis.

Draw a Conclusion:

Based on your decision, draw a conclusion in the context of the problem.

These tests provide a way to assess whether the observed differences in means are statistically significant or if they could have occurred by random chance. The appropriate test depends on the nature of your data and the study design.

PROPORTIONS AND EQUALITY OF PROPORTIONS

When you want to compare proportions or test the equality of proportions between two or more groups, you can use statistical tests designed for this purpose. Here are a couple of common tests:

Z-Test for Proportions (Two Independent Proportions):

Objective: To compare the proportions of two independent groups to determine if they are significantly different from each other.

Assumptions: The data should follow a binomial distribution, and the samples should be independent

Formula:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where:

- p_1 and p_2 are the proportions of the two groups.
- p is the combined proportion of the two groups.
- n_1 and n_2 are the sample sizes of the two groups.

Chi-Square Test for Independence (for Contingency Tables):

Objective: To test the association between two categorical variables, particularly when you want to assess if the proportions of one variable differ across different levels of another variable.

Assumptions: The data should be categorical and come from a random sample.

Formula: There isn't a direct formula for the Chi-Square test, as it involves constructing a contingency table and comparing the observed frequencies with the expected frequencies.

Steps for Hypothesis Testing:

Formulate Hypotheses:

H_0 : Null hypothesis (usually stating no difference in proportions or no association).1

H_1 or H_α : Alternative hypothesis (claiming a significant difference in proportions or an association).

Choose Significance Level (α):

Common choices are 0.05, 0.01, etc.

Collect and Analyze Data:

Collect your sample data.

Calculate the test statistic (z for z-test, chi-square for chi-square test).

Make a Decision:

If the p-value is less than or equal to α , reject the null hypothesis.

If the p-value is greater than α , fail to reject the null hypothesis.

Draw a Conclusion:

Based on your decision, draw a conclusion in the context of the problem.

These tests provide a way to assess whether the observed differences in proportions are statistically significant or if they could have occurred by random chance. The choice of the test depends on the nature of your data and the specific research question you are addressing.

MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY

CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatISTICS & DATA SCIENCE

SUBJECT CODE: 23PEBC23

SYLLABUS

UNIT: IV

UNIT II

Small sample tests – Students ‘t’ test for mean, difference of two way means, tests for correlation and regression coefficients. Chi-square test for goodness of a non-independence of attributes. F test for equality of variances. ANOVA- one way and two way. Basic concept related to biological studies.

Student's t-test for Mean:

Purpose:

Used to determine if there is a significant difference between the means of two independent groups.

Assumptions:

- Data within each group follows a normal distribution.
- Homogeneity of variances assumption (variances of the populations are equal).
- Independence of observations within and between groups.

Procedure:

1. Calculate the sample means (\bar{x}_1 and \bar{x}_2) for the two groups.

2. Calculate the sample standard deviations (s_1 and s_2) for the two groups.

3. Calculate the t-statistic using the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

4. Determine the degrees of freedom (df) using the formula:

$$df = n_1 + n_2 - 2$$

5. Compare the calculated t-value to the critical t-value from the t-distribution table at the desired significance level (α).

6. If the calculated t-value exceeds the critical t-value, reject the null hypothesis.

• Null Hypothesis (H_0):

- There is no significant difference between the means of the two groups.

• Alternative Hypothesis (H_a):

- There is a significant difference between the means of the two groups.

• Interpretation:

- If the p-value is less than the chosen significance level (α), typically 0.05, then the null hypothesis is rejected, suggesting that there is a significant difference between the means of the two groups.
- If the p-value is greater than α , then the null hypothesis is not rejected, suggesting that there is no significant difference between the means of the two groups.

• Example:

- Suppose we have two groups of students, Group A and Group B, and we want to determine if there is a significant difference in their mean test scores. We can use the t-test to compare the means of the two groups.

• Note:

- There are different variants of the t-test depending on the specific scenario, such as the paired t-test for dependent samples or the independent samples t-test for independent samples.

13. Chi-Square Test and Goodness of Fit

Chi-square test (χ^2) is applied in biostatistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is a measure to study the difference of actual and expected frequencies. It has great use in biostatistics especially in sampling studies.

In sampling studies, we never expect that there will be perfect coincidence between expected and observed frequencies. Since chi-square measures the difference between the expected and observed frequencies. If there is no difference between the actual and expected frequencies, χ^2 is zero. Thus, the chi-square test describes the discrepancy between theory and observation.

Characteristics of χ^2 Test

1. The test is based on events or frequencies and not based on mean or S.D, etc.
2. The test can be used between the entire set of observed and expected frequencies.
3. To draw inferences, this test is applied, especially testing the hypothesis.
4. It is a general test and is highly useful in research.

Assumptions

1. The observations must be large.
2. All the observations must be independent.
3. All the events must be mutually exclusive.
4. For comparison purposes, the data, must be in original units.

Degree of Freedom (df)

When we compare the computed value of χ^2 with the table value, the degree of freedom is evident. The degree of freedom means the number of classes to which values can be assigned. If we have n observed frequencies, the corresponding χ^2 distribution will have $(n-1)$ degrees of freedom. For example, in the case of tossing the coins, there are two possibilities or classes, namely *head* and *tail*. Here $df = n-1$ i.e. $n = \text{Head and tail} \therefore df = 2-1 = 1$. In such a way according to the classes we fix df , namely $n-1$.

Application of Chi-square Test

It is used to test the goodness of fit. The test enables to find out whether the difference between the expected and observed values is significant or not. If the difference is little then the fit is good, otherwise the fit is poor.

Definition

The χ^2 may be defined as

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] \quad \begin{array}{l} \text{where } O = \text{Observed frequencies} \\ E = \text{Expected frequencies} \\ \Sigma = \text{Sum of} \end{array}$$

Steps

1. A hypothesis is established i.e. Null hypothesis.
2. Calculate the difference between observed value and expected value $(O-E)$.
3. Square the deviations calculated $(O-E)^2$.
4. Divide the $(O-E)^2$ by its expected frequency $(O-E)^2/E$.

5. Add all the values obtained in step 4. $\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$

6. Find the chi-square from χ^2 table at certain level of significance, usually 5% or 1% level.

Inference

If the calculated value of χ^2 is greater than the table value of χ^2 at certain degree of level of significance, we reject the hypothesis. If the calculated value of χ^2 is zero, the observed values and expected values completely coincide. If the calculated value of χ^2 is less than table value at certain degree of level of significance, it is said to be non-significant. This implies that the difference between the observed and expected frequencies may be due to fluctuations in sampling.

Illustration - 1: A coin is tossed 100 times of which head comes 60 times and tail 40 times. Would you accept the hypothesis that the coin is normal having no bias for either head or tail.

Solution

Steps

1. Null hypothesis- i.e. the coin is normal having no bias for either head or tail.
2. Level of significance 5%.
3. Determining expected frequencies (E).

Possibilities	Observed frequencies (O)	Expected frequencies (E)
Head	60	50
Tail	40	50

4. Fixing the degrees of freedom $df = n-1$
 n = number of events or possibilities i.e. head and tail
 $n = 2-1 = 1$

5. Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = Observed value
E = Expected value

Possibilities	Observed frequency (O)	Expected frequency (E)	(O-E)	(O-E) ²	$\frac{(O-E)^2}{E}$
Head	60	50	60-50=10	(10) ² =100	$\frac{100}{50}=2.0$
Tail	40	50	40-50=-10	(-10) ² =100	$\frac{100}{50}=2.0$

$$= \sum \left[\frac{(O-E)^2}{E} \right] = 4.00$$

Calculated χ^2 value = 4.00

Table value at 5% level for one degree of freedom is 3.84.

Inference

The calculated χ^2 value (4.00) is greater than the table value (3.84). Therefore the hypothesis is rejected. In other words, the coin is defective with bias for head.

Illustration - 2: A dice is tossed 120 times with the following results:

No. turned up	1	2	3	4	5	6	Total
Frequency	30	25	18	10	22	15	120

Test the hypothesis that the dice is unbiased.

Solution

Steps

1. Null hypothesis - i.e. the dice is an unbiased one.
2. Level of significance 5%.

3. Determining expected frequencies (E).

The expected frequency is $[120 \times \frac{1}{6}] = 20$

4. Fixing the degrees of freedom

$$df = n-1$$

$$\text{i.e.} = 6-1 = 5$$

5. Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = Observed value

E = Expected value

No. turned up	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
1	30	20	10	100	5.00
2	25	20	5	25	1.25
3	18	20	-2	4	0.20
4	10	20	-10	100	5.00
5	22	20	2	4	0.20
6	15	20	-5	25	1.25
					$\Sigma = 12.90$

$$= \sum \left[\frac{(O-E)^2}{E} \right] = 12.90$$

Calculated χ^2 value = 12.90

For 5df, at 5% level of significance, the table value = 11.07

Inference

The calculated χ^2 value (12.90) is greater than the table value (11.07). Therefore the hypothesis is rejected. In other words, the dice is biased one.

Illustration - 3: A cross involving different genes gave rise to F_2 generation of tall and dwarf in the ratio of 110:90. Test by means of chi-square whether this value is deviated from the Mendel's monohybrid ratio 3:1.

Solution

Steps

1. **Null hypothesis:** There is no difference between 110:90 and Mendel's monohybrid ratio 3:1.

2. Level of significance 5%.

3. Determining expected frequencies (E).

Mendel's monohybrid ratio Tall: Dwarf = 3:1

Observed total number = 110+90 = 200

Expected = Tall and dwarf 3 : 1
= 150 : 50 = 200

4. Fixing the degrees of freedom

$$df = n-1$$

$$= 2-1 = 1$$

Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = Observed value

E = Expected value

Variables	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
Tall	110	150	-40	1600	10.6
Dwarf	90	50	40	1600	32.0
					$\Sigma = 42.6$

$$= \sum \left[\frac{(O-E)^2}{E} \right] = 42.6$$

Calculated χ^2 value = 42.6

For 1 df, at 5% level of significance the table value = 3.84

Inference

The calculated χ^2 value (42.6) is greater than the table value (3.84). Therefore the hypothesis is rejected. In other words the value 110:90 is deviated from Mendel's monohybrid ratio 3:1

Illustration - 4: When two heterozygous pea plants are crossed, 1600 plants are produced in the F_2 generation out of which 940 are yellow round, 260 are yellow wrinkled, 340 are green round and 60 are green wrinkled. By means of chi-square test whether these values are deviated from Mendel's dihybrid ratio 9 : 3 : 3 : 1. (or By means of chi-square test whether there is real independent assortment).

Solution

Steps

1. **Null/hypothesis:** There is real independent assortment i.e., there is no difference between observed values and Mendel's dihybrid ratio 9 : 3 : 3 : 1.

2. Level of significance 5%.

3. Determining expected frequencies (E) : Mendel's dihybrid ratio 9 : 3 : 3 : 1

$$\text{Yellow Round} = 9 \text{ Total } 1600 \therefore E = \frac{9}{16} \times 1600 = 900$$

$$\text{Yellow Wrinkled} = 3 \quad \therefore E = \frac{3}{16} \times 1600 = 300$$

$$\text{Green Round} = 3 \quad \therefore E = \frac{3}{16} \times 1600 = 300$$

$$\text{Green Wrinkled} = 1 \text{ Total } 1600 \therefore E = \frac{1}{16} \times 1600 = 100$$

4. Fixing the degrees of freedom $df = n-1$
 $= 4-1 = 3$

Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = Observed value
 E = Expected value

Variables	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
Yellow Round	940	900	40	1600	1.77
Yellow Wrinkled	260	300	-40	1600	5.33
Green Round	340	300	40	1600	5.33
Green Wrinkled	60	100	-40	1600	16.00
					$\Sigma = 28.43$

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right] = 28.43$$

Calculated χ^2 value = 28.43

For 3df, at 5% level of significance,

Table χ^2 value = 7.81

Inference

The calculated χ^2 value (28.43) is greater than the table χ^2 value (7.81). Therefore the hypothesis is rejected. In other words, there is no real independent assortment or the observed values are deviated from Mendel's dihybrid ratio 9 : 3 : 3 : 1.

Illustration - 5: When a black rat (heterozygous) is crossed with another heterozygous black rat, 43 black, 15 cream and 22 albino offspring are produced in the F_2 generation. Using chi-square, test the genetic hypothesis 9:3:4 is consistent with the data.

Solution

Steps

1. **Null hypothesis:** The genetic hypothesis 9 : 3 : 4 is consistent with the data.

2. Level of significance 5%.

3. Determining expected frequencies (E).

Genetic hypothesis = 9 : 3 : 4.

$$\therefore \text{Black} = 9 \quad \text{Total offspring} = 80 \therefore E = \frac{9}{16} \times 80 = 45$$

$$\text{Cream} = 3 \quad " = 80 \therefore E = \frac{3}{16} \times 80 = 15$$

$$\text{Albino} = \frac{4}{16} \quad " = 80 \therefore E = \frac{4}{16} \times 80 = \frac{20}{80}$$

4. Fixing the degrees of freedom $df = n-1$
 $= 3-1=2$

Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

where O = Observed value
 E = Expected value

Variables	O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
Black	43	45	-2	4	0.08
Cream	15	15	0	0	0
Albino	22	20	2	4	0.20
					$\Sigma = 0.28$

$$= \sum \left[\frac{(O-E)^2}{E} \right] = 0.28$$

Calculated χ^2 value = 0.28

For 2df, at 5% level of significance

the table χ^2 value = 5.96

Inference

The calculated χ^2 value (0.28) is less than the table χ^2 value (5.96). Therefore the hypothesis is accepted. In other words the observed value is consistent with the ratio 9 : 3 : 4.

Illustration - 6: A certain drug was administered to 500 people out of a total of 800 included in the sample to test its efficacy against typhoid. The results are given below: Find out the effectiveness of the drug against the disease (The table value of χ^2 for 1df at 5% level of significance is 3.84).

	Typhoid	No typhoid	Total
Administering the drug	200	300	500
Without administering the drug	280	20	300
Total	480	320	800

Solution Steps

1. Null hypothesis i.e. the drug is not effective in preventing typhoid.
2. Level of significance 5%.
3. Preparing 2x2 contingency table (observed value) [O].

	Typhoid	No typhoid	Total
Drug	200	300	500
No Drug	280	20	300
Total	480	320	800

4. Preparing table for expected frequencies (E).

	Typhoid	No Typhoid	Total
Drug	$\frac{480 \times 500}{800} = 300$	$\frac{320 \times 500}{800} = 200$	500
No. Drug	$\frac{480 \times 300}{800} = 180$	$\frac{320 \times 300}{800} = 120$	300
Total	480	320	800

N.B: Alternatively, after finding out the first value, the remaining values can be obtained easily in the following manner:

	Typhoid	No Typhoid	Total
Drug	$\frac{480 \times 500}{800} = 300$	$500 - 300 = 200$	500
No. Drug	$480 - 300 = 180$	$320 - 200 = 120$ [$300 - 180 = 120$]	300
Total	480	320	800

5. Fixing the degrees of freedom

$$\text{df} = (r-1)(c-1)$$

where r = row
 c = column

$$= (2-1)(2-1)$$

$$= 1$$

6. Calculation

$$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$$

O	E	O-E	(O-E) ²	$\frac{(O-E)^2}{E}$
200	300	-100	10000	33.33
280	180	100	10000	55.55
300	200	100	10000	50.00
20	120	-100	10000	83.33
				$\Sigma = 222.21$

$$= \sum \left[\frac{(O-E)^2}{E} \right] = 222.21$$

Calculated χ^2 value = 222.21

For 1 df, at 5% level of significance

the table χ^2 value = 3.84

Inference

The calculated χ^2 value (222.21) is greater than the table χ^2 value (3.84). Therefore the null hypothesis is rejected. In other words the drug is effective in preventing typhoid.



14. Analysis of Variance (ANOVA)

Analysis of variance refers to *the examination of differences among the samples*. It is an extremely useful technique concerning research in Biology. It is abbreviated as ANOVA (*Analysis of Variance* - AN + O + VA). It is used to examine the significance of the difference amongst more than two sample means at the same time. For example, when we want to compare more than two populations such as the yield of crop from several varieties of seeds, the drug habits of different groups of students and so on.

One can draw inferences about whether the samples have been drawn from population having the same mean, with the help of this technique.

The term *ANOVA* was first proposed by *R.A. Fisher*. He developed systematic procedure for the analysis of variation. It consists of classifying and cross-classifying statistical results and testing whether the means of specific classification differ significantly. For example, as mentioned above, five fertilizers are used to five plots of paddy. We may be interested in finding out whether the effect of these fertilizers on the yields are significantly different. To find out answer to this problem, we make use of ANOVA. It enables us to analyse the total variation into components which may be attributed to various '*sources*' or '*causes*'. It can provide us with meaningful comparisons of sample data which are classified according to two or more variables.

Assumptions in Analysis of Variance

1. Each of the sample is drawn from a normal population.
2. The variances for the population from which samples have been drawn are equal $\sigma_{12} = \sigma_{22} = \sigma_{32} = \sigma_{42}$ and so on.
3. The individuals being observed have been randomly, selected from the populations represented by samples.

Techniques of Analysis of Variance

The analysis of variance has been classified into

- a. *One-way classification* and
- b. *Two-way classification*

One-way Classification

In one-way classification, the data are classified according to only one criterion. That is, the arithmetic means of populations from which *K* samples were randomly drawn are equal to one another.

Principle

We take two estimates of population variance i.e., One based on *between samples variance* and the other *within samples variance*. Then these two estimates of population variance are compared with 'F' test as follows.

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

The value of F is to be compared to the F-limit for a given degrees of freedom. If the calculated F value exceeds the F-table value, we can say that there are significant variance between the sample means.

Steps Involved in the Analysis are:

Step : 1

Find out the means of each samples

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \dots, \bar{x}_k$$

Step : 2

Find out the combined mean of the samples

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4 + \dots + \bar{x}_k}{\text{No. of samples}}$$

Step : 3

Take the deviations of the sample means (Step 2) from mean of each sample. The square of such deviations which can be multiplied by the number of items in the corresponding samples. Then total the values. This is named as *sum of squares between the samples* (or) *SS - between*

$$\therefore \text{SS-between} = n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2 + \dots + n_k(\bar{x}_k - \bar{x})^2$$

n = number of items in the corresponding samples.

Step : 4

Divide the result of the 3rd step by the degrees of freedom. This is named as *Mean square between the samples* (or) *MS - between*

$$\therefore \text{Ms-between} = \frac{\text{SS-between}}{\text{degrees of freedom between the samples}}$$

Step : 5

Find out the deviations of the values of the sample items for all the samples from corresponding means of the samples. Then squares of such deviations and finally total the values. This is named as *sum of squares within the samples* (or) *SS-within*

$$\therefore \text{SS-within} = \sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)^2 + \sum (x_4 - \bar{x}_4)^2 + \dots + \sum (x_k - \bar{x}_k)^2$$

Step : 6

Divide the result of the 5th step by the degrees of freedom. This is named as *Mean square within the samples* (or) *MS-within*

$$\therefore \text{MS-within} = \frac{\text{SS - within}}{\text{degrees of freedom within the samples}}$$

Step : 7

Make ANOVA Table

Source of Variation	Sum of squares SS	Degree of freedom	Mean square MS
Between samples			
Within samples			
Total			

Step : 8

Find out F - value

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{\text{MS - between}}{\text{MS - within}}$$

If the calculated F-value is less than F-Table value, there is no significant.

Illustration - 1: A certain manure was used on four plots of land A, B, C and D. Four beds were prepared in each plot and the manure used. The output of the crop in the beds of plots A, B, C and D is given below:

A	B	C	D
6	15	9	8
8	10	3	12
10	4	7	1
8	7	1	3

Using ANOVA find out whether the difference in the means of the production of crops of the plots is significant or not

Solution

Step : 1

Find out the means of each sample.

	Sample I x_1	Sample II x_2	Sample III x_3	Sample IV x_4
	6	15	9	8
	8	10	3	12
	10	4	7	1
	8	7	1	3
Total	32	36	20	24
Average : \bar{x}	8	9	5	6

Step : 2

Find out the combined mean of the samples.

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \bar{x}_4}{\text{Number of samples}}$$

$$= \frac{8 + 9 + 5 + 6}{4}$$

$$= \frac{28}{4} = 7$$

$$\bar{x} = 7$$

Step : 3

Take the deviations of the sample means (Step 2) from mean of each sample. The square of such deviations which can be multiplied by the number of items in the corresponding samples. Then total the values. This is named as **sum of squares between the samples** (or) **SS-between**

$$\begin{aligned} \therefore \text{SS-between} &= n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2 + n_4(\bar{x}_4 - \bar{x})^2 \\ &= 4(8-7)^2 + 4(9-7)^2 + 4(5-7)^2 + 4(6-7)^2 \\ &= 4(1)^2 + 4(2)^2 + 4(-2)^2 + 4(-1)^2 \\ &= 4(1) + 4(4) + 4(4) + 4(1) \\ &= 4 + 16 + 16 + 4 \\ &= 40 \end{aligned}$$

Step : 4

Divide the result of the 3rd step by the degrees of freedom. This is named as **mean square between the samples** (or) **MS-between**.

$$\therefore \text{MS-between} = \frac{\text{SS-between}}{\text{degrees of freedom between the samples}}$$

There are four samples so the degrees of freedom are $4-1 = 3$

$$\therefore \text{MS-between} = \frac{40}{3} = 13.33$$

Step : 5

Find out the deviations of the values of the sample items for all the samples from corresponding means of the samples. Then squares of such deviations and finally total the values. This is named as **sum of squares within the samples** (or) **SS-within**.

$$\therefore \text{SS-within} = \sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \sum (x_3 - \bar{x}_3)^2 + \sum (x_4 - \bar{x}_4)^2$$

Sample I			Sample II		
x_1	$(x_1 - \bar{x}_1)$ $\bar{x}_1 = 8$	$(x_1 - \bar{x}_1)^2$	x_2	$(x_2 - \bar{x}_2)$ $\bar{x}_2 = 9$	$(x_2 - \bar{x}_2)^2$
6	$6-8 = -2$	4	15	$15-9 = 6$	36
8	$8-8 = 0$	0	10	$10-9 = 1$	1
10	$10-8 = 2$	4	4	$4-9 = -5$	25
8	$8-8 = 0$	0	7	$7-9 = -2$	4
$\sum (x_1 - \bar{x}_1)^2 =$		8	$\sum (x_2 - \bar{x}_2)^2 =$		66

Sample III			Sample IV		
x_3	$(x_3 - \bar{x}_3)$ $\bar{x}_3 = 5$	$(x_3 - \bar{x}_3)^2$	x_4	$(x_4 - \bar{x}_4)$ $\bar{x}_4 = 6$	$(x_4 - \bar{x}_4)^2$
9	$9 - 5 = 4$	16	8	$8 - 6 = 2$	4
3	$3 - 5 = -2$	4	12	$12 - 6 = 6$	36
7	$7 - 5 = 2$	4	1	$1 - 6 = -5$	25
1	$1 - 5 = -4$	16	3	$3 - 6 = -3$	9
$\Sigma(x_3 - \bar{x}_3)^2 =$		40	$\Sigma(x_4 - \bar{x}_4)^2 =$		74

$$\therefore \text{SS-within} = 8 + 66 + 40 + 74$$

$$= 188$$

Step : 6

Divide the result of the 5th step by the degrees of freedom. This is named as **Mean square within the samples** (or) **MS-within**

$$\therefore \text{MS-within} = \frac{\text{SS-within}}{\text{degrees of freedom within the samples}}$$

There are 16 items within the 4 samples.

$$\therefore \text{degrees of freedom } 16 - 4 = 12$$

$$\therefore \text{MS-within} = \frac{188}{12}$$

$$= 15.66$$

Step : 7

Make ANOVA Table

Source of variation	Sum of squares Ss	Degrees of freedom (Df)	Mean square (Ms)
Between samples	40	3	13.33
Within samples	188	12	15.66
Total	228	15	

Step : 8

Find out F value

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{13.33}{15.66} = 0.851$$

$$= 0.851$$

The table value of F for $v_1 = 3$ and $v_2 = 12$ at 5% level of significance = 3.49

Step : 9

Inference: The calculated value (0.851) is lesser than the table value (3.49). Therefore the difference in the means of the production of crops of the plots is not significant.

Short-cut Method (Indirect method)

Step : 1

Total sum of all the items of various samples (here 4 samples).

Sample I		Sample II		Sample III		Sample IV	
x_1	x_1^2	x_2	x_2^2	x_3	x_3^2	x_4	x_4^2
6	36	15	225	9	81	8	64
8	64	10	100	3	9	12	144
10	100	4	16	7	49	1	1
8	64	7	49	1	1	3	9
Σx_1 = 32	Σx_1^2 = 264	Σx_2 = 36	Σx_2^2 = 390	Σx_3 = 20	Σx_3^2 = 140	Σx_4 = 24	Σx_4^2 = 218

Total sum of all the items of various samples = T

$$T = \Sigma x_1 + \Sigma x_2 + \Sigma x_3 + \Sigma x_4$$

$$= 32 + 36 + 20 + 24 = 112$$

$$T = 112$$

Step : 2
 Correction factor = $\frac{T^2}{N}$ where N = number of items

$$= \frac{(112)^2}{16} = \frac{12544}{16} = 784$$

Step : 3
 Total sum of squares (or) **Total SS**
 Total SS = $[\sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2] - \text{correction factor}$

$$= [264 + 390 + 140 + 218] - 784$$

$$= 1012 - 784$$

$$= 228$$

Step : 4
 Sum of squares between samples (or) **SS-between**

$$\text{SS-between} = \frac{(\sum x_1)^2}{N} + \frac{(\sum x_2)^2}{N} + \frac{(\sum x_3)^2}{N} + \frac{(\sum x_4)^2}{N} - \text{correction factor}$$

$$= \frac{(32)^2}{4} + \frac{(36)^2}{4} + \frac{(20)^2}{4} + \frac{(24)^2}{4} - 784$$

$$= \frac{1024}{4} + \frac{1296}{4} + \frac{400}{4} + \frac{576}{4} - 784$$

$$= [256 + 324 + 100 + 144] - 784$$

$$= 824 - 784$$

$$= 40$$

Step : 5
 Sum of squares within samples (or) **SS-within**
 $\therefore \text{SS-within} = \text{Total sum of squares} - \text{Sum of squares between samples}$

$$= [\text{Total SS}] - [\text{SS-between}]$$

 (or)

$$= [\text{Value of step 3}] - [\text{Value of step 4}]$$

$$= [228] - [40]$$

$$= 228 - 40 = 188$$

$$= 188$$

Step : 6
 Make ANOVA Table

Source of variation	Sum of squares Ss	Degree of freedom (Df)	Mean square (Ms)
Between samples	40	3	$\frac{40}{3} = 13.33$
Within samples	188	12	$\frac{188}{12} = 15.66$
Total	228	15	

Step:7

Find out F value

$$F = \frac{\text{Variance between samples}}{\text{Variance within samples}} = \frac{13.33}{15.66} = 0.851$$

$$= 0.851$$

The table value of F for $v_1 = 3$ and $v_2 = 12$ at 5% level of significance = 3.49

Step : 8

Inference: The calculated value (0.851) is lesser than the table value (3.49). Therefore the difference in the means of the production of crops of the plots is not significant.

Analysis of Variance in Two-way Classification: (Two way ANOVA)

It is used when the data are classified on the basis of two factors. For example, the yield of rice, in addition to being affected by fertility of land, might also be affected by monsoon, quality of seeds, fertilizers, modern technique, etc. when it is believed that two independent factors have an affect in the response variable, an analysis of variance can be used to test for the effects of the two factors simultaneously. Such a test is called *two way ANOVA* (or) *two factor analysis of variance*.

In a two-way classification the data are classified according to two different criteria. Make two-way ANOVA table.

Sources of Variation	Sum of squares SS	Degree of freedom (Df)	Mean squares MS	F
Between columns	SSC	$c - 1$	$MSC = \frac{SSC}{c - 1}$	$\frac{MSC}{MSE}$
Between rows	SSR	$r - 1$	$MSR = \frac{SSR}{r - 1}$	$\frac{MSR}{MSE}$
Residual error	SSE	$(c-1)(r-1)$	$MSE = \frac{SSE}{(c-1)(r-1)}$	
Total	SST	$n-1$		

Where, SSC - Sum of squares between columns

SSR - Sum of squares between rows

SSE - Sum of squares due to error

MSC - Mean square between columns

MSR - Mean square between rows

MSE - Mean square due to error

SST - Total sum of squares

The sum of squares for the source 'Residual' (SSE) is obtained by subtracting from the total sum of squares (SST) by the sum of squares between columns (SSC) and rows (SSR) i.e., $SSE = SST - (SSC + SSR)$.

Illustration - 2: Set up two-way ANOVA table for the following results.

Per acre production data for sorghum.

Name of fertilizers	Variety of Sorghum seeds		
	Co.1	Co.5	Co.9
Urea	6	5	5
Ammonium sulphate	7	5	4
Zinc sulphate	3	3	3
Potash	8	7	4

Solution

Step : 1

Calculate Total = T and the number of items = N

$$T = 6+7+3+8+5+5+3+7+5+4+3+4 = 60$$

$$N = 12$$

Step : 2

$$\text{Correction Factor} = T^2/N = \frac{(60)^2}{12} = \frac{3600}{12}$$

$$= 300$$

Step : 3

Square of all items

$$= (6)^2 + (7)^2 + (3)^2 + (8)^2 + (5)^2 + (5)^2 + (3)^2 + (7)^2 + (5)^2 + (4)^2 + (3)^2 + (4)^2$$

$$= 36 + 49 + 9 + 64 + 25 + 25 + 9 + 49 + 25 + 16 + 9 + 16$$

$$= 332$$

Step : 4

Total sum of squares (SST)

$$SST = \text{Square of all items} - \text{Correction Factor}$$

$$\text{i.e.} = [\text{Value of step 3} - \text{Correction Factor}]$$

$$= 332 - 300 = 32$$

$$SST = 32$$

Step : 5

Sum of squares between variety of sorghum seeds (or)

$$\text{SS between columns} = \frac{(\sum \text{Co.1})^2}{N} + \frac{(\sum \text{Co.5})^2}{N} + \frac{(\sum \text{Co.9})^2}{N} - \text{Correction factor}$$

Name of fertilizers	Variety of Sorghum seeds			
	Co.1	Co.5	Co.9	Total
Urea x_1	6	5	5	$16 = \sum x_1$
Ammonium sulphate x_2	7	5	4	$16 = \sum x_2$
Zinc sulphate x_3	3	3	3	$9 = \sum x_3$
Potash x_4	8	7	4	$19 = \sum x_4$
Total	$\sum Co.1 = 24$	$\sum Co.5 = 20$	$\sum Co.9 = 16$	60

\therefore SS between columns

$$\begin{aligned}
 &= \frac{(24)^2}{4} + \frac{(20)^2}{4} + \frac{(16)^2}{4} - \text{Correction factor} \\
 &= \left[\frac{576}{4} + \frac{400}{4} + \frac{256}{4} \right] - 300 \\
 &= [144 + 100 + 64] - 300 \\
 &= 8
 \end{aligned}$$

Step : 6

Sum of squares between fertilizers (or)

$$\begin{aligned}
 \text{SS between rows} &= \frac{(\sum x_1)^2}{N} + \frac{(\sum x_2)^2}{N} + \frac{(\sum x_3)^2}{N} + \frac{(\sum x_4)^2}{N} - \text{Correction factor} \\
 &= \frac{(16)^2}{3} + \frac{(16)^2}{3} + \frac{(9)^2}{3} + \frac{(19)^2}{3} - \text{Correction factor} \\
 &= \frac{256}{3} + \frac{256}{3} + \frac{81}{3} + \frac{361}{3} - 300 \\
 &= [85.3 + 85.3 + 27 + 120.3] - 300 \\
 &= 317.9 - 300 \\
 &= 17.9 \\
 &= 18
 \end{aligned}$$

Step : 7

SS for error (SSE)

$$\text{SSE} = \text{SST} - (\text{SSC} + \text{SSR})$$

= Total sum of squares - (Sum of squares between columns + sum of squares between rows)
(In short)

SS for error

$$\begin{aligned}
 (\text{SSE}) &= \text{Total SS} - (\text{SS between columns} + \text{SS between rows}) \\
 &= [\text{Value of step 4}] - [\text{Value of step 5} + \text{Value of step 6}] \\
 &= 32 - (8 + 18) \\
 &= 32 - 26 \\
 &= 6
 \end{aligned}$$

Step : 8

Degrees of freedom

$$\begin{aligned}
 \text{d.f. for total variance} &= (c \times r - 1) \\
 &= (3 \times 4 - 1) \\
 &= 12 - 1 \\
 &= 11 \\
 \text{d.f. for variance between columns} &= (c - 1) \\
 &= 3 - 1 \\
 &= 2 \\
 \text{d.f. for variance between rows} &= (r - 1) \\
 &= 4 - 1 \\
 &= 3 \\
 \text{d.f. for residual variance} &= (c - 1)(r - 1) \\
 &= (3 - 1)(4 - 1) \\
 &= 2 \times 3 \\
 &= 6
 \end{aligned}$$

Step : 9

$$\begin{aligned}
 \text{MS between columns (MSC)} &= \frac{\text{SS between columns}}{c - 1} \\
 \text{i.e., MSC} &= \frac{\text{SSC}}{c - 1}
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\text{Value of step 5}}{c-1} \\
 &= \frac{8}{3-1} = \frac{8}{2} \\
 &= 4 \\
 \text{MS between rows (MSR)} &= \frac{\text{SS between rows}}{r-1} \\
 \text{i.e., MSR} &= \frac{\text{SSR}}{(r-1)} \\
 &= \frac{\text{Value of step 6}}{r-1} \\
 &= \frac{18}{4-1} = \frac{18}{3} \\
 &= 6 \\
 \text{MS residual or error (MSE)} &= \frac{\text{SS residual or error}}{(c-1)(r-1)} \\
 \text{i.e., MSE} &= \frac{\text{SSE}}{(c-1)(r-1)} \\
 &= \frac{\text{Value of step 7}}{(c-1)(r-1)} \\
 &= \frac{6}{(3-1)(4-1)} = \frac{6}{2 \times 3} \\
 &= \frac{6}{6} \\
 &= 1
 \end{aligned}$$

Step : 10**Setting two-way ANOVA Table**

Sources of variation	Sum of squares Ss	D.f	Mean square Ms	F Calculated Value	F- Table Value at 5%
Between columns	8 (SSC)	(c-1) 2	8/2 = 4 (MSC)	$F = \frac{\text{MSC}}{\text{MSE}} = \frac{4}{1} = 4$	5.14
Between rows	18 (SSR)	(r-1) 3	18/3 = 6 (MSR)	$F = \frac{\text{MSR}}{\text{MSE}} = \frac{6}{1} = 6$	4.76
SS residual or error	6 (SSE)	(c-1)(r-1) 6	6/6 = 1 (MSE)		
Total	32	11			

Step : 11

Inference: Since the F-ratio concerning the varieties of sorghum seeds (4.0) is less than table value (5.14). Therefore the differences concerning the varieties of sorghum seeds are *insignificant* at 5%.

But the differences concerning fertilizers are *significant* at 5% because the calculated value - F (6.0) is more than the table value (4.76).



Table - 6: Probability Distribution of X^2

n	0.99	0.98	0.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.0157	.0628	0.004	.0158	.0642	.148	.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	.0201	.0404	0.103	.211	.446	.713	1.386	2.408	3.219	4.605	5.991	7.824	9.120	13.815
3	0.115	.185	0.352	.584	1.005	1.421	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	0.297	.429	0.711	1.065	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	0.554	.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.107	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	5.812	6.614	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	29.633	32.00	39.252
17	6.408	7.255	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.790
18	7.015	7.906	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.989	28.869	32.346	34.805	42.312
19	7.633	8.567	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191	43.820
20	8.260	9.237	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566	45.315

21	8.897	9.915	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932	46.797
22	9.542	10.600	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.183	33.924	37.659	40.289	48.268
23	10.196	11.293	13.091	14.848	17.187	19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638	49.728
24	10.856	11.992	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980	51.179
25	11.524	12.697	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314	52.620
26	12.198	13.409	15.379	17.292	19.820	21.792	25.366	29.246	31.795	35.563	38.885	42.856	45.642	54.052
27	12.879	14.125	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963	55.476
28	13.565	14.847	16.928	18.939	21.588	23.647	27.336	31.391	34.027	37.916	41.337	45.419	48.278	56.893
29	14.256	15.574	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.693	49.588	58.302
30	14.953	16.306	18.493	20.599	23.364	25.508	29.336	33.530	36.250	40.256	43.773	47.962	50.892	59.703
32	16.362	17.783	20.072	22.271	25.148	27.373	31.336	35.665	38.466	42.585	46.194	50.487	53.486	62.487
34	17.789	19.275	21.664	23.952	26.938	29.242	33.336	37.795	40.676	44.903	48.602	52.995	56.061	65.247
36	19.233	20.783	23.269	25.643	28.735	31.115	35.336	39.922	42.879	47.212	50.999	55.489	58.619	67.985
38	20.691	22.304	24.884	27.343	30.537	32.992	37.335	42.045	45.076	49.513	53.384	57.969	61.162	70.703
40	22.164	23.838	26.509	29.051	32.345	34.872	39.333	44.165	47.269	51.805	55.759	60.436	63.691	73.402
42	23.650	25.383	28.144	30.765	34.157	36.755	41.335	46.282	49.456	54.090	58.124	62.892	66.206	76.084
44	25.148	26.939	29.787	32.487	35.974	38.641	43.335	48.396	51.639	56.369	60.481	65.337	68.710	78.750
46	26.657	28.504	31.439	34.215	37.795	40.529	45.335	50.507	53.818	58.641	62.830	67.771	71.201	81.400
48	28.177	30.080	33.098	35.949	39.621	42.420	47.335	52.616	55.993	60.907	65.171	70.197	73.683	84.037
50	29.707	31.664	34.764	37.689	41.449	44.313	49.335	54.723	58.164	63.167	67.505	72.613	76.154	86.661
52	31.246	33.256	36.437	39.433	43.281	46.209	51.335	56.827	60.332	65.422	69.832	75.021	78.616	89.272
54	32.793	34.856	38.116	41.183	45.117	48.106	53.335	58.930	62.496	67.673	72.153	77.422	81.069	91.872
56	34.350	36.464	39.801	42.937	46.955	50.005	55.335	61.031	64.658	69.919	74.468	79.815	83.513	94.461
58	35.913	38.078	41.492	44.696	48.797	51.906	57.335	63.129	66.816	72.160	76.778	82.201	85.950	97.039
60	37.485	39.699	43.188	46.459	50.641	53.809	59.335	65.227	68.972	74.397	79.082	84.580	88.379	99.607

Table - 7: Table of F (Variance Ratio) 1% Points

$\frac{N^1}{N^2}$	1	2	3	4	5	6	8	12	24	α
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98.49	99.01	99.17	99.25	96.30	99.33	99.36	99.42	99.46	99.50
3	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.12
4	21.20	18.00	15.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.27	9.89	9.47	9.02
6	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96	3.59	3.16
14	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80	3.43	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.39	4.77	4.44	4.20	3.89	3.55	3.18	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45	3.08	2.65
18	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.79	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.76	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.81
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
α	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Table - 8: Table of F -5 % Points

N_1 N_2	1	2	3	4	5	6	8	12	24	α
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	253.4
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.83	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.64	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
24	4.26	3.40	3.10	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.89	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.74	2.52	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
α	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00



MARUDHAR KESARI JAIN COLLEGE FOR WOMEN, VANIYAMBADI

PG AND RESEARCH DEPARTMENT OF BIOCHEMISTRY

CLASS: I M.SC BIOCHEMISTRY

SUBJECT NAME: BIostatistics & Data Science

SUBJECT CODE: 23PEBC23

SYLLABUS

UNIT: V

UNIT V

Introduction to Data Science, Definition of data science, importance, and basic applications, Machine Learning Algorithms, Deep Learning, Artificial Neural Networks and their Application, Reinforcement Learning, Natural Language Processing Artificial Intelligence (AI), Data Visualization, Data Analysis, Optimization Techniques, Big Data, Predictive Analysis. Application of AI in medical, health and pharma industries.

Introduction to Data Science, Definition of data science

Data science is a multidisciplinary field that encompasses various techniques, methodologies, and tools aimed at extracting actionable insights and knowledge from data. It involves the application of statistics, machine learning, computer science, domain expertise, and other related disciplines to analyze large volumes of structured and unstructured data.

At its core, data science involves the following key components:

- **Data Acquisition:** Gathering data from various sources, including databases, sensors, social media, and other platforms.
- **Data Cleaning and Preparation:** Processing and transforming raw data into a structured format suitable for analysis. This step often involves handling missing values, removing duplicates, and standardizing data.
- **Exploratory Data Analysis (EDA):** Exploring and visualizing the data to understand its underlying patterns, trends, and relationships.
- **Statistical Analysis:** Applying statistical techniques to derive insights and make inferences from the data.
- **Machine Learning:** Utilizing algorithms and models to uncover complex patterns, make predictions, and automate decision-making processes.
- **Data Interpretation and Communication:** Interpreting the results of analyses and communicating findings to stakeholders through reports, dashboards, and visualizations.
- Data science is used across various industries and domains, including but not limited to finance, healthcare, retail, marketing, and technology. It plays a crucial role in informing strategic decision-making, optimizing processes, improving customer experiences, and driving innovation.

Overall, data science enables organizations to leverage their data assets effectively, gain valuable insights, and stay competitive in today's data-driven world.

The importance of data science

The importance of data science stems from its ability to unlock insights and drive decision-making processes across various domains. Here are some key reasons why data science is important:

Informed Decision Making: Data science empowers organizations to make data-driven decisions based on evidence and insights derived from large volumes of data. This leads to more accurate predictions, better strategies, and optimized outcomes.

Improved Efficiency: By automating processes, identifying inefficiencies, and optimizing workflows, data science can significantly improve operational efficiency and reduce costs for businesses.

Enhanced Customer Experiences: Data science enables organizations to better understand customer behavior, preferences, and needs through techniques such as customer segmentation and sentiment analysis. This insight can be used to personalize products, services, and marketing efforts, leading to enhanced customer satisfaction and loyalty.

Predictive Analytics: Through predictive modeling and forecasting techniques, data science allows businesses to anticipate future trends, behaviors, and events. This helps organizations to proactively address potential challenges, seize opportunities, and stay ahead of the competition.

Risk Management: Data science plays a crucial role in identifying and mitigating risks across various sectors, including finance, healthcare, and cybersecurity. By analyzing historical data and identifying patterns, organizations can better assess and manage risks, ultimately improving resilience and stability.

Innovation and Research: Data science drives innovation by enabling researchers and scientists to analyze large datasets, uncover new insights, and make breakthrough discoveries. This is particularly evident in fields such as healthcare, where data science is used to accelerate drug discovery, personalize treatments, and improve patient outcomes.

Applications of data science

Some basic applications of data science include:

Recommendation Systems: Used in e-commerce, streaming platforms, and social media to personalize recommendations based on user preferences and behavior.

Fraud Detection: Employed in finance, insurance, and cybersecurity to detect and prevent fraudulent activities by analyzing patterns and anomalies in transactional data.

Healthcare Analytics: Utilized to improve patient outcomes, optimize hospital operations, and develop personalized treatment plans based on medical history and genomic data.

Supply Chain Optimization: Applied in logistics and manufacturing to optimize inventory management, streamline distribution processes, and reduce costs.

Sentiment Analysis: Used in marketing and social media to analyze customer sentiment, opinions, and trends, enabling organizations to understand public perception and tailor their messaging accordingly.

Overall, the importance and applications of data science continue to grow as organizations recognize the value of leveraging data to drive innovation, improve decision-making, and achieve strategic objectives.

Machine learning algorithms

Machine learning algorithms are a crucial component of data science, enabling computers to learn from data and make predictions or decisions without being explicitly programmed. These algorithms are categorized into several types based on their learning style and application.

Here are some common types of machine learning algorithms:

Supervised Learning:

Classification: Used for predicting categorical labels or classes. Examples include logistic regression, decision trees, random forests, support vector machines (SVM), and k-nearest neighbors (KNN).

Regression: Used for predicting continuous numerical values. Examples include linear regression, polynomial regression, and ridge regression.

Unsupervised Learning:

Clustering: Used for grouping similar data points together based on their characteristics. Examples include k-means clustering, hierarchical clustering, and DBSCAN.

Dimensionality Reduction: Used for reducing the number of features in a dataset while preserving its essential information. Examples include principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE).

Association Rule Learning: Used for discovering interesting relationships or associations among variables in large datasets. Examples include Apriori algorithm and FP-growth algorithm.

Semi-supervised Learning:

Combines elements of both supervised and unsupervised learning. It leverages a small amount of labeled data along with a larger amount of unlabeled data to improve model performance.

Reinforcement Learning:

Involves an agent learning to make decisions by interacting with an environment. The agent receives feedback in the form of rewards or penalties based on its actions. Examples include Q-learning, deep Q-networks (DQN), and policy gradients.

Deep Learning:

Deep learning is a subset of machine learning that focuses on training artificial neural networks with multiple layers (hence the term "deep") to learn representations of data. These neural networks are composed of interconnected nodes, called neurons, organized into layers. Each layer processes the data in a hierarchical manner, with higher layers learning increasingly abstract features from the input data.

Key concepts and components of deep learning include:

Artificial Neural Networks (ANNs): These are computational models inspired by the structure and functioning of the human brain. ANNs consist of interconnected nodes organized into layers, including an input layer, one or more hidden layers, and an output layer. Each connection between nodes has an associated weight that is adjusted during the training process to minimize the error between the predicted output and the actual output.

Deep Neural Networks (DNNs): DNNs are neural networks with multiple hidden layers, allowing them to learn complex representations of data. Deep learning architectures can range from relatively shallow networks with a few hidden layers to very deep networks with dozens or even hundreds of layers.

Convolutional Neural Networks (CNNs): CNNs are a type of deep neural network designed for processing grid-like data, such as images and videos. They consist of convolutional layers, pooling layers, and fully connected layers. CNNs leverage parameter sharing and local connectivity to extract spatial hierarchies of features from the input data, making them highly effective for tasks such as image recognition and object detection.

Recurrent Neural Networks (RNNs): RNNs are specialized neural networks designed for sequential data, such as time series or natural language sequences. Unlike feedforward neural networks, RNNs have connections that form directed cycles, allowing them to capture temporal dependencies in the data. Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are popular variants of RNNs that address the vanishing gradient problem and facilitate learning long-term dependencies.

Generative Adversarial Networks (GANs): GANs are a class of deep learning models that consist of two neural networks, a generator and a discriminator, which are trained simultaneously in a competitive fashion. The generator aims to generate realistic samples from a given distribution, while the discriminator learns to distinguish between real and generated samples. GANs have applications in generating synthetic data, image synthesis, and unsupervised representation learning.

Deep learning has achieved remarkable success in various domains, including computer vision, natural language processing, speech recognition, and reinforcement learning. Its ability to automatically learn hierarchical representations of data from raw inputs has led to significant advances in artificial intelligence and has enabled groundbreaking applications in fields such as healthcare, finance, autonomous vehicles, and more.

Artificial Neural Networks and its applications

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functioning of the human brain. They consist of interconnected nodes, called neurons or artificial neurons, organized into layers. Each neuron receives input signals, processes them using an activation function, and produces an output signal that is passed to neurons in the next layer.

ANNs are widely used in various applications across different domains due to their ability to learn complex patterns and make predictions from data. Some common applications of artificial neural networks include:

Image Recognition and Computer Vision:

Convolutional Neural Networks (CNNs) are a specialized type of ANN widely used for image recognition tasks, such as object detection, image classification, and facial recognition. CNNs have

achieved remarkable accuracy in tasks like identifying objects in images, detecting anomalies in medical images, and recognizing handwritten characters.

Natural Language Processing (NLP):

Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) are commonly used in NLP tasks, such as text classification, sentiment analysis, machine translation, and speech recognition. These networks can model sequential dependencies in text data, making them effective for tasks that involve analyzing and generating natural language.

Predictive Analytics and Time Series Forecasting:

ANNs, including feedforward neural networks and recurrent neural networks, are used for predictive analytics tasks, such as financial forecasting, stock price prediction, demand forecasting, and weather prediction. These networks can learn patterns from historical data and make predictions about future trends or events.

Healthcare and Medical Diagnosis:

Artificial neural networks are applied in healthcare for tasks such as disease diagnosis, medical image analysis, drug discovery, and personalized medicine. CNNs are used for medical image analysis tasks like tumor detection in MRI scans, while RNNs are used for time-series data analysis in patient monitoring and disease progression prediction.

Autonomous Vehicles and Robotics:

Neural networks are employed in autonomous vehicles for tasks like object detection, lane detection, path planning, and decision-making. CNNs are used for processing sensor data, such as images and LiDAR scans, to detect and classify objects in the vehicle's environment.

Recommendation Systems:

Neural networks are used in recommendation systems to personalize recommendations for users based on their preferences and behavior. Collaborative filtering and deep learning-based approaches, such as neural collaborative filtering and deep recommendation models, are commonly used in recommendation systems for e-commerce, streaming platforms, and content recommendation.

These are just a few examples of the wide-ranging applications of artificial neural networks. As the field of deep learning continues to advance, neural networks are being applied to increasingly complex tasks across various domains, driving innovation and transforming industries.

Reinforcement learning

Reinforcement learning (RL) is a type of machine learning paradigm where an agent learns to make decisions by interacting with an environment. The agent takes actions in the environment, receives feedback in the form of rewards or penalties, and learns to maximize its cumulative reward over time. Unlike supervised learning, where the model learns from labeled data, and unsupervised

learning, where the model learns from unlabeled data, reinforcement learning deals with learning from a series of actions and their consequences.

Key components and concepts of reinforcement learning include:

Agent: The entity that learns and makes decisions. It interacts with the environment by observing its current state, selecting actions, and receiving rewards or penalties based on its actions.

Environment: The external system or process with which the agent interacts. The environment responds to the actions taken by the agent and transitions to new states based on those actions.

State: A representation of the current situation or configuration of the environment. The state provides information about the context in which the agent is making decisions.

Action: The choices available to the agent at each time step. The agent selects actions based on its current state and the policy it follows.

Reward: A scalar feedback signal received by the agent after taking an action in a particular state. The reward indicates the immediate desirability of the action and serves as feedback for learning.

Policy: A strategy or set of rules that governs the agent's behavior. The policy determines the mapping from states to actions and guides the agent's decision-making process.

Value Function: A function that estimates the expected cumulative reward of being in a particular state or taking a particular action. Value functions help the agent evaluate the desirability of different states or actions.

Exploration vs. Exploitation: The trade-off between exploring new actions and exploiting known actions with high expected rewards. RL algorithms balance exploration and exploitation to learn optimal policies efficiently.

Popular algorithms and techniques in reinforcement learning include:

Q-Learning: A model-free RL algorithm that learns an action-value function (Q-function) to estimate the expected cumulative reward of taking an action in a particular state.

Deep Q-Networks (DQN): Extends Q-learning by using deep neural networks to approximate the Q-function, enabling RL in environments with large state spaces.

Policy Gradient Methods: Directly optimize the policy function to maximize expected rewards. Examples include REINFORCE, Actor-Critic methods, and Proximal Policy Optimization (PPO).

Temporal Difference Learning: Update value estimates based on the difference between predicted and observed rewards, incorporating information from successive time steps.

Reinforcement learning has applications in a wide range of domains, including robotics, autonomous systems, gaming, finance, healthcare, and recommendation systems. It is used to develop agents that can learn to play games, control robots, navigate complex environments, optimize resource allocation, and make decisions in dynamic and uncertain environments. RL

continues to be an active area of research, driving innovations in artificial intelligence and autonomous systems.

Natural Language Processing

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that focuses on enabling computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP encompasses a wide range of tasks and techniques for processing and analyzing natural language data, including text and speech.

Key components and concepts of natural language processing include:

- **Tokenization:** The process of breaking down a piece of text into smaller units, such as words or sentences, called tokens. Tokenization is the first step in many NLP tasks.
- **Part-of-Speech Tagging:** Assigning grammatical categories (e.g., noun, verb, adjective) to each word in a sentence based on its syntactic role.
- **Named Entity Recognition (NER):** Identifying and classifying named entities, such as persons, organizations, locations, dates, and numerical expressions, in a text.
- **Parsing:** Analyzing the grammatical structure of a sentence to determine its syntactic relationships, such as subject-verb-object.
- **Sentiment Analysis:** Determining the sentiment or emotional tone of a piece of text, such as positive, negative, or neutral.
- **Topic Modeling:** Discovering the underlying themes or topics present in a collection of documents.
- **Machine Translation:** Translating text from one language to another automatically using computational methods.
- **Text Generation:** Generating new text based on a given input or context, such as auto-completion, language modeling, and dialogue systems.
- **Question Answering:** Automatically answering questions posed in natural language based on a given context or knowledge base.
- **Information Extraction:** Extracting structured information from unstructured text, such as extracting entities, relations, and events from news articles or documents.
- **Text Classification:** Assigning predefined categories or labels to text documents based on their content, such as spam detection, sentiment classification, and topic classification.
- **Word Embeddings:** Representing words or phrases as dense vectors in a continuous vector space, capturing semantic similarities and relationships between words.
- Popular tools and libraries for natural language processing include NLTK (Natural Language Toolkit), spaCy, Stanford NLP, Transformers (using libraries like Hugging Face), and Gensim.

NLP has a wide range of applications across various domains, including:

- Search engines
- Virtual assistants (e.g., Siri, Alexa)
- Chatbots and conversational agents
- Social media analysis

- Customer feedback analysis
- Sentiment analysis in product reviews
- Text summarization
- Language translation
- Information retrieval
- Medical text analysis
- Legal document analysis

Overall, natural language processing plays a crucial role in enabling machines to understand and interact with human language, driving innovation in many areas of technology and industry.

Artificial Intelligence (AI)

Artificial Intelligence (AI) refers to the development of computer systems or machines that can perform tasks that typically require human intelligence. These tasks include understanding natural language, recognizing patterns, learning from experience, reasoning, problem-solving, and adapting to new situations. AI systems aim to mimic cognitive functions associated with human intelligence, such as perception, learning, reasoning, and decision-making.

Key components and techniques within artificial intelligence include:

Machine Learning: Machine learning is a subset of AI that focuses on building systems capable of learning from data and making predictions or decisions without being explicitly programmed. It encompasses various algorithms and approaches, including supervised learning, unsupervised learning, reinforcement learning, and deep learning.

Natural Language Processing (NLP): NLP involves the interaction between computers and human language. It encompasses tasks such as text parsing, sentiment analysis, language translation, text summarization, and question answering. NLP techniques enable machines to understand, interpret, and generate human language, facilitating communication between humans and computers.

Computer Vision: Computer vision is the field of AI concerned with enabling machines to interpret and understand visual information from the surrounding environment. It involves tasks such as object detection, image classification, image segmentation, facial recognition, and scene understanding. Computer vision algorithms allow machines to perceive and interpret visual data, enabling applications in autonomous vehicles, surveillance, medical imaging, and augmented reality.

Robotics: Robotics combines AI with mechanical engineering to create autonomous or semi-autonomous machines capable of performing physical tasks in various environments. AI algorithms enable robots to perceive their surroundings, plan and execute actions, and interact with objects and humans in real-world scenarios. Robotics applications include industrial automation, healthcare assistance, household chores, and exploration of hazardous environments.

Knowledge Representation and Reasoning: AI systems often rely on formal representations of knowledge to facilitate reasoning and decision-making. Knowledge representation techniques

involve encoding knowledge in a structured format that machines can understand and manipulate. This knowledge can be used to infer new facts, solve problems, and make informed decisions.

Expert Systems: Expert systems are AI systems designed to mimic the decision-making abilities of human experts in specific domains. They incorporate knowledge bases, inference engines, and reasoning algorithms to analyze input data, provide recommendations, and solve complex problems. Expert systems have applications in fields such as healthcare, finance, engineering, and customer support.

Cognitive Computing: Cognitive computing refers to AI systems that aim to simulate human thought processes, such as perception, reasoning, and learning. These systems leverage techniques from AI, machine learning, natural language processing, and other fields to emulate human-like intelligence. Cognitive computing applications include virtual assistants, personalized recommendations, and intelligent tutoring systems.

Overall, artificial intelligence encompasses a wide range of technologies, techniques, and applications aimed at creating intelligent systems capable of performing tasks that traditionally require human intelligence. AI continues to advance rapidly, driving innovation across various industries and reshaping the way we interact with technology.

Data Visualization, Data Analysis, and Optimization Techniques

Data Visualization, Data Analysis, and Optimization Techniques are closely related concepts within the realm of data science and analytics. Let's briefly explore each one:

Data Visualization:

Data visualization is the graphical representation of data and information to communicate insights, patterns, and trends effectively. It involves creating visual representations such as charts, graphs, maps, and dashboards to convey complex data in a clear and intuitive manner.

Data visualization helps analysts and decision-makers explore data, identify patterns, detect outliers, and communicate findings to stakeholders. It enables storytelling with data, allowing users to convey narratives and insights visually.

Popular data visualization tools and libraries include Tableau, ggplot2 (in R), Matplotlib and Seaborn (in Python), D3.js, Power BI, and Plotly.

Data Analysis:

Data analysis involves the process of inspecting, cleaning, transforming, and modeling data to extract meaningful insights and inform decision-making. It encompasses various techniques and methodologies, including descriptive statistics, inferential statistics, machine learning, and data mining.

Data analysis aims to uncover patterns, relationships, and trends within data, answer specific questions, and derive actionable insights to support business objectives or research goals.

Key steps in data analysis include data preprocessing, exploratory data analysis (EDA), hypothesis testing, model building and evaluation, and interpretation of results.

Optimization Techniques:

Optimization techniques are methods used to find the best solution to a problem within a given set of constraints. These techniques are applied to optimize processes, systems, and decisions to achieve desired outcomes efficiently.

Optimization problems can be classified into two main types: linear optimization (e.g., linear programming) and nonlinear optimization (e.g., gradient-based methods, genetic algorithms).

Optimization techniques are widely used in various domains, including operations research, logistics, supply chain management, finance, engineering design, and machine learning.

Common optimization algorithms and techniques include linear programming, integer programming, dynamic programming, gradient descent, genetic algorithms, simulated annealing, and convex optimization.

In practice, data visualization, data analysis, and optimization techniques often complement each other in the data science workflow. Data visualization helps analysts explore and understand data, data analysis techniques are used to extract insights and patterns, and optimization techniques are applied to improve processes, models, and decision-making based on those insights. Together, these concepts form a powerful toolkit for extracting value from data and driving informed decision-making in various domains.

Big data and predictive analysis

Big data and predictive analysis are two interconnected concepts in the field of data science and analytics. Let's explore each one:

Big Data:

Big data refers to extremely large and complex datasets that cannot be easily processed or analyzed using traditional data processing techniques. These datasets typically exhibit the three V's: volume (large amount of data), velocity (high speed of data generation), and variety (diverse types of data).

Big data sources include structured data from databases, unstructured data from social media, sensor data from IoT devices, text data from emails and documents, and multimedia data from images and videos.

Big data technologies and frameworks, such as Apache Hadoop, Apache Spark, and NoSQL databases, are used to store, process, and analyze large volumes of data efficiently and in parallel.

Big data analytics involves extracting insights, patterns, and trends from big data to support decision-making, optimize processes, and drive innovation across various domains.

Predictive Analysis:

Predictive analysis is the process of using historical data, statistical algorithms, and machine learning techniques to make predictions about future events or outcomes. It involves identifying patterns and relationships in data to forecast future trends, behaviors, or events.

Predictive analysis is applied in various domains, including finance (e.g., stock price prediction), marketing (e.g., customer churn prediction), healthcare (e.g., disease diagnosis and prognosis), and manufacturing (e.g., predictive maintenance).

Common predictive analysis techniques include regression analysis, time series forecasting, classification and regression trees, neural networks, and ensemble methods such as random forests and gradient boosting.

Predictive analysis enables organizations to anticipate future outcomes, mitigate risks, identify opportunities, and make informed decisions based on data-driven insights.

Big data and predictive analysis often go hand in hand, as big data provides the massive datasets needed to train predictive models and extract meaningful insights. Predictive analysis can leverage big data technologies and techniques to process and analyze large volumes of data efficiently, uncovering valuable patterns and relationships that can be used for forecasting and decision-making purposes. By combining big data and predictive analysis, organizations can unlock the potential of their data assets and gain a competitive advantage in today's data-driven world.

Artificial Intelligence (AI) is revolutionizing the medical, health, and pharmaceutical industries

Artificial Intelligence (AI) is revolutionizing the medical, health, and pharmaceutical industries by enhancing diagnosis, treatment, drug discovery, personalized medicine, and administrative processes. Here are some key applications of AI in these sectors:

Medical Imaging Analysis:

AI algorithms, particularly deep learning-based approaches, are used to analyze medical images such as X-rays, MRI scans, CT scans, and histopathology slides. These algorithms can assist radiologists and pathologists in detecting abnormalities, diagnosing diseases (e.g., cancer, Alzheimer's), and segmenting organs or tumors accurately.

Disease Diagnosis and Prognosis:

AI-powered diagnostic tools can analyze patient data, including medical history, symptoms, and test results, to assist healthcare providers in diagnosing diseases and predicting patient outcomes. These tools can improve diagnostic accuracy, reduce errors, and enable early detection of diseases.

Drug Discovery and Development:

AI is used in drug discovery and development to accelerate the identification of potential drug candidates, predict drug efficacy and safety, and optimize drug design. AI-driven approaches, such

as virtual screening, molecular modeling, and predictive analytics, help pharmaceutical companies streamline the drug discovery process and bring new treatments to market faster.

Personalized Medicine:

AI enables the development of personalized medicine approaches tailored to individual patients based on their genetic makeup, medical history, and other relevant factors. AI algorithms analyze large-scale genomic and clinical data to identify biomarkers, predict treatment responses, and recommend personalized treatment plans for patients with complex diseases.

Healthcare Robotics and Assistive Technologies:

AI-powered robots and assistive technologies are used in healthcare settings to assist with tasks such as surgery, patient monitoring, rehabilitation, and elderly care. Surgical robots can enhance precision and minimize invasiveness in procedures, while robotic exoskeletons and prosthetics can improve mobility and quality of life for patients with disabilities.

Clinical Decision Support Systems (CDSS):

AI-based CDSS provide healthcare providers with real-time clinical insights, evidence-based recommendations, and treatment guidelines to support decision-making at the point of care. These systems analyze patient data, medical literature, and best practices to assist clinicians in making informed decisions about diagnosis, treatment, and patient management.

Healthcare Administration and Operations:

AI is used to optimize healthcare administration and operations by automating administrative tasks, streamlining workflows, and improving resource allocation. AI-powered solutions can enhance patient scheduling, billing, electronic health record (EHR) management, supply chain management, and fraud detection, leading to cost savings and operational efficiencies.

Overall, AI has the potential to transform the medical, health, and pharmaceutical industries by improving patient outcomes, reducing healthcare costs, and enabling more efficient and personalized healthcare delivery. As AI technologies continue to advance, they are expected to play an increasingly important role in shaping the future of healthcare.

BIOSTATISTICS & DATA SCIENCE

UNIT I

2 MARK QUESTIONS

1. Define the nature of biological experiments.
2. Differentiate between primary and secondary data.
3. Define data.
4. Define Classification
5. Define tabulation in organizing data
6. Name two different forms of diagrams or graphs commonly used in representing biological data.
7. Define mean,
8. Define median,
9. Define mode.

5 MARK QUESTIONS

1. Discuss the role of primary data in biological and clinical experiments.
2. Compare and contrast the advantages and disadvantages of using secondary data in biological studies.
3. Describe three different methods commonly employed for collecting data in biological and clinical experiments.
4. Explain the steps involved in the process of classification and tabulation of data
5. Explore the diverse forms of diagrams and graphs used in biological studies.
6. Define mean, median, and mode as measures of averages.
7. Discuss the importance of statistical measures, such as mean, median, and mode, in drawing conclusions from experimental data.
8. How do mean, median, and mode contribute to data interpretation and presentation in biological studies?

10 MARK QUESTIONS

1. Discuss the challenges and advantages of using primary data in biological and clinical experiments.
2. Evaluate the ethical considerations associated with the use of secondary data in biological studies.
3. Compare and contrast the quantitative and qualitative methods commonly employed for data collection in biological and clinical experiments.
4. Elaborate on the importance of proper classification and tabulation of data in biological experiments.
5. Explore the principles of designing effective diagrams and graphs for biological studies.

QUESTION BANK

BIOSTATISTICS & DATA SCIENCE

UNIT I

2 MARK QUESTIONS

1. Define the nature of biological experiments.
2. Differentiate between primary and secondary data.
3. Define data.
4. Define Classification
5. Define tabulation in organizing data
6. Name two different forms of diagrams or graphs commonly used in representing biological data.
7. Define mean,
8. Define median,
9. Define mode.

5 MARK QUESTIONS

1. Discuss the role of primary data in biological and clinical experiments.
2. Compare and contrast the advantages and disadvantages of using secondary data in biological studies.
3. Describe three different methods commonly employed for collecting data in biological and clinical experiments.
4. Explain the steps involved in the process of classification and tabulation of data
5. Explore the diverse forms of diagrams and graphs used in biological studies.
6. Define mean, median, and mode as measures of averages.
7. Discuss the importance of statistical measures, such as mean, median, and mode, in drawing conclusions from experimental data.
8. How do mean, median, and mode contribute to data interpretation and presentation in biological studies?

10 MARK QUESTIONS

1. Discuss the challenges and advantages of using primary data in biological and clinical experiments.
2. Evaluate the ethical considerations associated with the use of secondary data in biological studies.
3. Compare and contrast the quantitative and qualitative methods commonly employed for data collection in biological and clinical experiments.
4. Elaborate on the importance of proper classification and tabulation of data in biological experiments.
5. Explore the principles of designing effective diagrams and graphs for biological studies.

6. Analyze the statistical measures of mean, median, and mode in the context of biological studies. Discuss their respective strengths and limitations and how researchers can choose the most appropriate measure based on the nature of their data.
7. Discuss the role of averages (mean, median, mode) in hypothesis testing and data interpretation in biological experiments.
8. Evaluate the importance of statistical software and tools in analyzing complex biological data.
9. Explore the potential biases and limitations associated with averages in biological studies.
10. Discuss the integration of various statistical measures and data visualization techniques in the presentation of experimental results in biological studies.

UNIT II

2 MARK QUESTIONS

1. Define quartile deviation.
2. Define mean deviation.
3. Define standard deviation.
4. Define the coefficient of variation.
5. Define skewness and kurtosis
6. Briefly provide relationship between correlation and regression
7. Define rank correlation.
8. Define a regression equation?

5 MARK QUESTIONS

1. Discuss the applications of quartile deviation, mean deviation, and standard deviation in analyzing the variability of biological characters.
2. Explain the significance of the coefficient of variation in comparing the variability of different biological characters.
3. Describe the characteristics of a positively skewed distribution in biological data.
4. Discuss the implications of kurtosis in the distribution of biological characters.
5. Compare and contrast correlation and regression in the context of biological studies. Provide examples
6. Explain the process of rank correlation and its advantages in situations where data may not follow a normal distribution.
7. Define a regression equation and discuss its practical application in predicting one biological variable based on another. Provide a real-world example to illustrate.
8. Explore the challenges and benefits of using statistical measures in analyzing biochemical data.

10 MARK QUESTIONS

1. Evaluate the strengths and limitations of quartile deviation, mean deviation, and standard deviation as measures of dispersion for biological characters.
2. Discuss the applications of the coefficient of variation in comparing the variability of biological characters across different experimental conditions.

3. Analyze the impact of skewness on the interpretation of experimental results in biological studies.
4. Explore the role of kurtosis in the analysis of frequency distributions for biological characters.
5. Compare and contrast the assumptions, applications, and limitations of correlation and regression in analyzing biological data.
6. Discuss the significance of rank correlation in situations where normal distribution assumptions may not hold for biological data.
7. Explain the steps involved in deriving and interpreting a regression equation for biological characters.
8. Evaluate the importance of statistical measures in analyzing complex biochemical data.
9. How do measures of dispersion, correlation, and regression contribute to a comprehensive understanding of the biochemical characteristics under investigation?
10. Discuss the ethical considerations in the collection and analysis of biological data.
11. How can researchers ensure the integrity and validity of their statistical findings in biological studies?
12. Explore the practical challenges and strategies for handling outliers in statistical analyses of biological data.
13. How do outliers impact measures of dispersion, correlation, and regression, and what methods can be employed to address them?

UNIT III

2 MARK QUESTIONS

1. Define a simple random sample.
2. Give the concept of a stratified sample
3. Define systematic sampling
4. Define sampling distribution.
5. Define standard error.
6. Explain the rationale behind conducting a test of significance based on large samples.
7. Define a test for the mean
8. Explain the concept of a test for the difference of means
9. Define a test for proportions

5 MARK QUESTIONS

1. Discuss the advantages and limitations of simple random sampling in the context of biological research.
2. Compare and contrast stratified sampling and systematic sampling.
3. Explain the concept of sampling distribution.
4. Define standard error and discuss its significance in statistical inference.
5. Discuss the rationale behind conducting a test of significance based on large samples.
6. Describe the steps involved in a test for the mean in biological research.
7. Explore the concept of a test for the difference of means.
8. Define a test for proportions and discuss its application in analyzing categorical data.

9. Explain the purpose of a test for the equality of proportions.

10 MARK QUESTIONS

1. Evaluate the strengths and weaknesses of simple random sampling as a method of data collection in biological research.
2. Compare the advantages and disadvantages of stratified sampling and systematic sampling.
3. Discuss the theoretical underpinnings of the sampling distribution.
4. Explain the concept of standard error and its role in estimating the precision of sample statistics.
5. Explore the considerations and implications of conducting a test of significance based on large samples in biological research.
6. Describe the steps involved in conducting a test for the mean in biological research. Discuss potential challenges and considerations in the interpretation of results.
7. Analyze the application of a test for the difference of means in comparing two groups in a biological experiment.
8. Define a test for proportions and discuss its utility in analyzing categorical data. Provide examples of situations where this test is particularly relevant in biological research.
9. Explain the purpose and significance of a test for the equality of proportions.
10. Discuss the practical implications and challenges associated with choosing an appropriate sampling method, conducting statistical tests, and interpreting results in a real-world biological study.

UNIT IV

2 MARK QUESTIONS

1. What is the primary purpose of the Student's t-test?
2. Provide a brief explanation of the Chi-square test.
3. Define the F test.
4. Define one-way ANOVA.
5. Define correlation coefficients in biological studies.
6. Define the purpose of tests for regression coefficients and their importance in regression analysis in biological research.
7. In the context of statistical testing,
8. what does the p-value indicate?
9. Differentiate between the t-test for mean and the t-test for the difference of two means.
10. Define two-way ANOVA
11. Define the term "null hypothesis"

5 MARK QUESTIONS

1. Discuss the key assumptions and limitations associated with the Student's t-test for mean.

2. Explain the process of conducting a Chi-square test for goodness of non-independence of attributes.
3. Compare and contrast the F test for equality of variances with the Levene's test.
4. Elaborate on the steps involved in performing one-way ANOVA.
5. Discuss the interpretation of correlation coefficients in the context of biological research.
6. Explore the significance of tests for regression coefficients in regression analysis.
7. Describe the concept of statistical power in hypothesis testing.
8. Differentiate between Type I and Type II errors in the context of hypothesis testing.
9. Explain the role of interaction effects in two-way ANOVA.
10. Discuss the ethical considerations related to statistical hypothesis testing in biological research.

10 MARK QUESTIONS

1. Evaluate the strengths and weaknesses of the Student's t-test for mean in the context of biological research.
2. Discuss the practical steps and considerations in performing a Chi-square test for goodness of non-independence of attributes in a biological study.
3. Compare and contrast different methods for assessing equality of variances, including the F test and Levene's test.
4. Discuss their respective advantages and limitations in biological experiments.
5. Explore the assumptions underlying one-way ANOVA.
6. How do violations of these assumptions affect the reliability of ANOVA results in biological studies?
7. Elaborate on the interpretation of correlation coefficients, considering both magnitude and direction.
8. Discuss the applications of tests for regression coefficients in regression analysis.
9. Explain the concept of statistical power in hypothesis testing.
10. Differentiate between Type I and Type II errors in hypothesis testing.
11. Analyze the complexities of interaction effects in two-way ANOVA.
12. Discuss the ethical considerations associated with statistical hypothesis testing in biological research.

UNIT V

2 MARK QUESTIONS

1. Define data science.
2. Discuss the importance of data science.
3. Provide two examples of basic applications of data science.
4. Define machine learning algorithms
5. Explain the concept of deep learning.
6. Define artificial neural networks.
7. Briefly explain reinforcement learning
8. Define natural language processing
9. Provide a concise definition of artificial intelligence.

10. Define data visualization in data science
11. Define data analysis
12. Define optimization techniques in the context of data science.
13. Define big data.
14. Define predictive analysis
15. Give the application of artificial intelligence

5 MARK QUESTIONS

1. Discuss the role of machine learning algorithms in improving decision-making processes in businesses.
2. Explain the significance of artificial neural networks in handling complex data patterns.
3. Describe the fundamental principles of reinforcement learning and its application in training intelligent systems.
4. Explore the challenges and opportunities associated with the application of natural language processing (NLP) in understanding and interpreting human language in data science.
5. Discuss the importance of data visualization in conveying meaningful insights.
6. Explain the concept of big data and discuss how it differs from traditional data.
7. Explore the applications of artificial intelligence in the medical, health, and pharmaceutical industries.
8. Discuss the role of optimization techniques in data science.
9. Describe the basic principles of predictive analysis and discuss its significance in anticipating future trends and outcomes in various domains.
10. Explain the concept of data analysis in the context of data science.

10 MARK QUESTIONS

1. Evaluate the impact of machine learning algorithms on decision-making processes in businesses.
2. Explore the architecture and functioning of artificial neural networks.
3. Discuss in detail the principles of reinforcement learning and its application in training intelligent systems.
4. Analyze the ethical considerations associated with the application of natural language processing (NLP) in data science.
5. Investigate the challenges and benefits of handling big data in data science.
6. Assess the role of artificial intelligence in revolutionizing the medical, health, and pharmaceutical industries.
7. Explore optimization techniques in the context of data science.
8. Discuss the evolution and applications of predictive analysis in various industries.
9. Describe the significance of data analysis in data science.
10. Evaluate the challenges and opportunities associated with integrating artificial intelligence, machine learning, and big data in comprehensive data science projects.

Find the standard deviation in individual observation of following data ?

60, 61, 60, 63, 61, 63, 62, 64, 70, 64

Calculate the standard deviation for the following data (Direct method)

Size of the item (x)	6	7	8	9	10	11	12
frequency	3	6	9	13	8	5	4

Calculate the regression analysis for the following data of yield of tomato and potato. What is the probable yield of potato, when the yield of tomato happens to be 150 kg.

Tomato (x)	60	20	10	40	80	150
Potato (y)	90	60	50	80	120	?