XI Computer Science 2025

Slides Revision: 1.0

Data and Analysis

S.L.O. # 4

Sub Topics: 3 Total SLO: 10

MCQ: (3) 3 Marks CRQ: (1) 3 Marks ERQ: (0) 0 Marks

Prepared By: Naseer Ahmad (Nusrat Jahan Boys College) 0333-9790903

4.1 Statistical Modelling

SLO	Students should be able to	Cognitive Level
4.1.1	define the following terms: a. statistics, b. statistical modeling;	R
4.1.2	explain the following statistical modeling techniques (supervised and unsupervised): a. regression, b. classification, c. k-means clustering;	U
4.1.3	analyse the process of developing a statistical model in the context of data analysis;	An
4.1.4	evaluate the effectiveness of statistical modeling in solving real-world problems in the fields of healthcare and finance;	E

SLO 4.1.1 R

define the following terms:

a. statistics,
b. statistical modeling;

Statistics an Statistical Modeling

SLO 4.1.1 R

a. **Statistics**: Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data to make informed decisions or draw conclusions about a population based on a sample.

b. **Statistical Modeling**: Statistical modeling is the process of creating mathematical representations (models) that describe relationships between variables in data, enabling prediction, inference, or understanding of underlying phenomena.

SLO 4.1.2 U

explain the following statistical modeling techniques (supervised and unsupervised):

a. regression,
b. classification,
c. k-means clustering;

Supervised vs Unsupervised Learning

SLO 4.1.2 U

Feature	Supervised Learning	Unsupervised Learning
What data looks like	Every example has a **label** (correct answer)	No labels at all – just raw data
Real-life analogy	A teacher giving you questions **with answers** to study	You're given thousands of photos and told "figure out what groups exist" without any names
Goal	Learn to predict the label for new, unseen data	Discover hidden patterns or groups in the data
Types of problems	Regression (predict a number) Classification (predict a category)	Clustering Dimensionality reduction Anomaly detection

Supervised Statistical Modeling Techniques

SLO 4.1.2 U

a. **Regression**:

- Purpose: Predict a numeric outcome (e.g., house price, temperature, sales revenue).
- Input: Features (independent variables) + known continuous target values (during training).
- Output: A continuous predicted value.
- Common methods:
  - Linear Regression
  - Polynomial Regression
  - Ridge / Lasso / Elastic Net (regularized linear regression)
  - Support Vector Regression (SVR)
  - Regression Trees, Random Forests, Gradient Boosting (XGBoost, LightGBM, CatBoost)
  - Neural Networks for regression

Supervised Statistical Modeling Techniques

SLO 4.1.2 U

a. **Regression Examples**:

- Predicting house prices in your city
  - A real estate company collects data on 10,000 sold houses: size (sqft), number of bedrooms, age of house, distance to city center, and the actual sale price (target). The regression model learns the relationship (e.g., +$300 per extra sqft, –$50,000 if the house is >30 years old). After training, you input the features of a new house and it predicts the sale price (e.g., $587,200).
- Predicting a person's blood pressure
  - Doctors measure age, BMI, salt intake, exercise hours/week, and record the actual systolic blood pressure for thousands of patients. A regression model learns the pattern and can now estimate blood pressure for a new patient before measuring it.

Supervised Statistical Modeling Techniques

SLO 4.1.2 U

b. **Classification**:

- Purpose: Assign data points to discrete classes or categories (e.g., spam/not spam, disease/no disease, cat/dog/ bird).
- Input: Features + known class labels (during training).
- Output: Predicted class (or probability of belonging to each class).
- Common methods:
  - Logistic Regression (binary or multinomial)
  - Decision Trees, Random Forests, Gradient Boosting
  - Support Vector Machines (SVM)
  - k-Nearest Neighbors (k-NN)
  - Naive Bayes
  - Neural Networks (including deep learning classifiers)

Supervised Statistical Modeling Techniques

SLO 4.1.2 U

b. **Classification Examples**:

- Email spam filter (Gmail, Outlook)
  - Millions of emails are labeled by users as “Spam” or “Not Spam.” The classification model (often logistic regression or a deep neural network) learns which words, sender patterns, and links are typical of spam. When a new email arrives, it instantly predicts: Spam (move to spam folder) or Not Spam.
- Bank deciding whether to approve a credit-card application
  - The bank has historical data: income, credit score, age, number of late payments → and whether the person defaulted (Yes/No). A classification model learns the pattern and predicts for a new applicant: “Approve” or “Reject.”

Supervised Statistical Modeling Techniques

SLO 4.1.2 U

b. Classification Examples:

- Airport security facial recognition
  - System trained on thousands of passenger photos labeled with their identity (or “authorized” vs “not authorized”). When you step up to the e-gate, the classification model decides in <1 second: “Match – open gate” or “No match – alert officer.”

Unsupervised Statistical Modeling Technique

SLO 4.1.2 U

c. **K-means Clustering**:

- Purpose: Partition data into k groups where points within the same cluster are more similar to each other than to points in other clusters.
- Input: Only features (no labels or target variable).
- Output: Assignment of each data point to one of k clusters (and cluster centroids).
- Objective: Minimize the within-cluster sum of squares (inertia).
- Important notes:
  - Assumes clusters are spherical and roughly equal in size.
  - Sensitive to initialization and the choice of k (use elbow method, silhouette score, etc., to choose k).
  - Hard assignment (each point belongs to exactly one cluster).

Unsupervised Statistical Modeling Technique

SLO 4.1.2 U

c. **K-means Clustering Examples**:

- Netflix or Spotify creating user "taste groups" without being told what the groups are
  - They take millions of users and look only at listening/viewing behavior (how many action movies, jazz songs, comedies, etc.).
  - k-means (or similar clustering) automatically discovers, say, 20 clusters such as "80s rock lovers," "true-crime documentary fans," "classical music seniors."
  - Netflix then recommends movies that people in your discovered cluster liked.
  - No one had to label the clusters in advance.

Unsupervised Statistical Modeling Technique

SLO 4.1.2 U

c. **K-means Clustering Examples**:

- Customer segmentation for a supermarket chain
  - The store has purchase data (frequency, amount spent, types of products bought) for 500,000 loyalty-card holders but no predefined segments.
  - k-means finds natural groups such as:
    - Budget shoppers buying mostly discounts
    - Health-conscious organic buyers
    - Families buying lots of baby products
    - Weekend wine & cheese buyers.
  - The marketing team then sends targeted coupons to each group

SLO 4.1.3 An

analyse the process of developing a statistical model in the context of data analysis;

SLO 4.1.3 An

Analysis of the Process of Developing a Statistical Model in Data Analysis

Developing a statistical model is a systematic process aimed at understanding data patterns, making predictions, or drawing inferences. The process typically involves the following key steps:

1. **Problem Definition**: Clearly define the objective of the analysis. This includes identifying the question to be answered, the type of outcome (continuous or categorical), and the context of the problem.
   Example: Predict the selling price of houses based on their features.
2. **Data Collection**: Gather relevant and sufficient data from reliable sources. Data quality is critical, as noisy or incomplete data can adversely affect model performance.
   Example: Collect data on houses including size, number of rooms, location, age, and sale price from real estate listings.
3. **Data Cleaning and Preprocessing**: Prepare the data by handling missing values, removing outliers, transforming variables, and encoding categorical data if necessary. This step ensures the data is in a usable form.
   Example: Remove entries with missing prices, fix inconsistent location names, and convert categorical location data into numerical format.

SLO 4.1.3 An

Analysis of the Process of Developing a Statistical Model in Data Analysis

4. **Exploratory Data Analysis (EDA):**Analyze the data using statistical summaries and visualization techniques to understand underlying patterns, distributions, and relationships between variables. EDA helps in selecting relevant features and informs model choice.
   Example: Plot scatter plots of house size vs price to see if larger houses generally sell for more, or check correlations between features.
5. **Feature Selection and Engineering**: Select significant variables that influence the outcome and engineer new features if needed to improve model accuracy and interpretability.
   Example: Select features like size, number of bedrooms, and neighborhood quality. Create a new feature such as "age category" (new, mid-age, old).
6. **Choosing a Modeling Technique**: Based on the problem type (regression, classification, clustering) and data characteristics, select an appropriate statistical modeling technique.
   Example: Use linear regression to predict continuous house prices.

SLO 4.1.3 An

Analysis of the Process of Developing a Statistical Model in Data Analysis

7. **Model Training**: Fit the chosen model to the training data, estimating parameters that best capture the relationship between input variables and the outcome.
   Example: Use the training dataset to estimate coefficients that relate house size and other features to price.
8. **Model Evaluation**: Assess the model's performance using appropriate metrics (e.g., mean squared error for regression, accuracy or F1 score for classification) on validation data to ensure it generalizes well to unseen data.
   Example: Calculate the mean squared error (MSE) on a test set to check prediction accuracy.
9. **Model Tuning and Optimization**: Adjust model parameters or hyperparameters to improve accuracy and prevent overfitting or underfitting.
   Example: Try adding polynomial terms or regularization to reduce overfitting and improve predictions.

SLO 4.1.3 An

Analysis of the Process of Developing a Statistical Model in Data Analysis

7. **Model Interpretation**: Interpret the model results to provide meaningful insights or actionable conclusions relevant to the original problem.
   Example: The model shows that each additional bedroom increases house price by $20,000 on average.
8. **Deployment and Monitoring**: Deploy the model for practical use and monitor its performance over time to maintain accuracy and relevance.
   Example: Implement the model in a real estate app to estimate prices, and monitor prediction accuracy monthly to update as market trends change.

SLO 4.1.4 E

evaluate the effectiveness of statistical modeling in solving real-world problems in the fields of healthcare and finance;

SLO 4.1.4 E

Evaluation of Statistical Modeling Effectiveness in Healthcare and Finance

Healthcare

- **Effectiveness**:
  - Disease Diagnosis and Prediction: Statistical models help predict disease risk based on patient data, improving early diagnosis. For example, logistic regression models predict the likelihood of diabetes or heart disease by analyzing factors like age, weight, and blood pressure.
  - Personalized Treatment: Models analyze patient responses to treatments, enabling personalized medicine. Predictive models can determine which drug or therapy is likely to be most effective for an individual.
  - Resource Optimization: Statistical forecasting models assist hospitals in predicting patient admissions and managing resource allocation efficiently.
- **Challenges**:
  - Data quality and privacy concerns can limit model accuracy.
  - Complex biological systems sometimes produce nonlinear interactions that simple models struggle to capture.
  - Ethical concerns arise when models affect critical health decisions.

SLO 4.1.4 E

Evaluation of Statistical Modeling Effectiveness in Healthcare and Finance

Finance

- **Effectiveness**:
  - Risk Assessment: Regression and classification models predict creditworthiness, enabling lenders to assess borrower risk accurately.
  - Fraud Detection: Unsupervised models like clustering detect unusual transaction patterns indicating fraud, improving security.
  - Algorithmic Trading: Statistical models analyze historical market data to identify trading opportunities, automating decisions at high speed.
- **Challenges**:
  - Financial markets are highly volatile, making predictions uncertain.
  - Models can be overfitted to past data, failing during unexpected market conditions.
  - Regulatory requirements impose constraints on model transparency and explainability.

4.2 Experimental Design in Data Science

SLO	Students should be able to	Cognitive Level
4.2.1	define the following terms: a. correlation, b. causation, c. population, d. parameters, e. random sample;	R
4.2.2	differentiate between observational studies and experimental studies in data science;	U
4.2.3	describe the principles of experimental design flow;	U
4.2.4	explain the steps involved in the experimental design flow;	U

Prepared By: Naseer Ahmad (Nusrat Jahan Boys College) 0333-9790903

SLO 4.2.1 R

define the following terms:

a. correlation,
b. causation,
c. population,
d. parameters,
e. random sample;

Define the Terms

SLO 4.2.1 R

- **Correlation**: A statistical measure that indicates the strength and direction of a linear relationship between two variables. It shows how variables move together but does not imply that one causes the other.
- **Causation**: A relationship where one event or variable directly causes a change in another. Unlike correlation, causation implies a cause-and-effect link.
- **Population**: The entire set of individuals or items that are the subject of a statistical study.
- **Parameters**: Numerical characteristics or measures that describe a population, such as the population mean or population standard deviation.
- **Random Sample**: A subset of a population selected in such a way that every member of the population has an equal chance of being chosen, ensuring unbiased representation.

SLO 4.2.2 U

differentiate between observational studies and experimental studies in data science;

SLO 4.2.2 U

Observational Studies vs Experimental Studies

Aspect	Observational Studies	Experimental Studies
Definition	Researchers observe and analyze data without intervening or manipulating variables.	Researchers actively intervene by applying treatments or changes to study subjects to observe effects.
Control over Variables	No control; variables occur naturally.	High control; researchers assign treatments or conditions.
Purpose	To identify associations or correlations.	To establish causal relationships.
Example	Studying the relationship between smoking and lung cancer by surveying smokers and non-smokers.	Conducting a clinical trial to test the effect of a new drug on patient health.

Observational Studies vs Experimental Studies

SLO 4.2.2 U

Aspect	Observational Studies	Experimental Studies
Randomization	Usually absent; subjects are observed as they are.	Typically involves random assignment of subjects to treatment or control groups.
Strengths	Easier to conduct, useful when experiments are unethical or impractical.	Stronger evidence of causality, controlled environment reduces confounding factors.
Limitations	Cannot definitively establish causation due to potential confounding variables.	Can be expensive, time-consuming, and sometimes ethically challenging.

SLO 4.2.3 U

describe the principles of experimental design flow;

Principles of Experimental Design Flow

SLO 4.2.3 U

1. Define the Objective: Clearly state the purpose of the experiment and the hypotheses to be tested.
2. Select Factors and Levels: Identify the independent variables (factors) to manipulate and decide the specific values or settings (levels) for each factor.
3. Choose the Experimental Units: Decide the subjects or items on which the experiment will be conducted (e.g., patients, machines, plots of land).
4. Randomization: Randomly assign experimental units to different treatment groups to reduce bias and evenly distribute unknown confounding factors.
5. Replication: Repeat the experiment or treatment conditions on multiple experimental units to measure variability and improve the precision of results.

Principles of Experimental Design Flow

SLO 4.2.3 U

6. Control: Use control groups or baseline treatments to compare effects and isolate the impact of the factors under study.
7. Blocking (if needed):Group similar experimental units together to control variability caused by nuisance e factors, allowing clearer assessment of treatment effects.
8. Conduct the Experiment: Implement the treatments as per the design, ensuring consistent conditions and accurate data collection.
9. Data Collection and Analysis: Collect results systematically, then analyze using appropriate statistical methods to test hypotheses.
10. Interpretation and Conclusion: Draw conclusions based on the analysis, considering the design, limitations, and real-world implications.

SLO 4.2.4 U

explain the steps involved in the experimental design flow;

Steps Involved in Experimental Design Flow

SLO 4.2.4 U

1. Identify the Research Problem and Objectives: Begin by clearly defining the question or problem the experiment aims to address. Formulate specific hypotheses or goals.
2. Select the Factors, Levels, and Responses:
   - Factors: Decide which independent variables (factors) will be manipulated.
   - Levels: Determine the different values or settings for each factor.
   - Response Variables: Decide what outcomes or responses will be measured to evaluate the effects.
3. Choose Experimental Units: Identify the subjects or items (people, plants, machines) on which the experiment will be performed.
4. Design the Experiment (Treatment Allocation):Plan how treatments (combinations of factor levels) will be assigned to experimental units. This includes deciding the number of treatments and their arrangement.
5. Randomization: Randomly assign experimental units to treatment groups to avoid systematic bias and ensure the results are generalizable.

Steps Involved in Experimental Design Flow

SLO 4.2.4 U

6. Replication: Repeat treatments on multiple experimental units to measure natural variability and increase reliability of results.
7. Control and Blocking (if necessary): Use control groups or standard conditions to compare against treatment effects. Apply blocking to group similar units and reduce variation from external factors.
8. Conduct the Experiment: Execute the experiment carefully, maintaining consistent conditions and accurately applying treatments.
9. Collect Data: Systematically record the results or responses from each experimental unit.
10. Analyze Data: Use statistical tools and methods (like ANOVA, regression) to examine the effects of treatments and test hypotheses.
11. Draw Conclusions: Interpret the analysis results in the context of the research objectives, discussing implications, limitations, and possible next steps.

4.3 Statistics and Data Visuals

SLO	Students should be able to	Cognitive Level
4.3.1	explain the data collection methods and its types;	U
4.3.2	analyse the role of data science in addressing real-world problems, supported by practical examples.	An

SLO 4.3.1 U

explain the data collection methods and its types;

Data Collection Methods and Its Types

SLO 4.3.1 U

Data Collection Methods

- Data collection is the process of gathering information from various sources to address research questions or test hypotheses. Effective data collection ensures accuracy, reliability, and relevance of the data for analysis.

Data Collection Methods and Its Types

SLO 4.3.1 U

Types of Data Collection Methods

1. Primary Data Collection: Data collected directly by the researcher for a specific purpose. This method provides original, firsthand information.
   - Surveys and Questionnaires: Structured tools with predefined questions to collect responses from participants.
     Example: Customer satisfaction surveys.
   - Interviews: Direct, often face-to-face, interactions to gather detailed information. Can be structured, semi-structured, or unstructured.
     Example: In-depth interviews with experts.
   - Observations: Systematic recording of behaviors or events as they naturally occur without interference.
     Example: Monitoring customer behavior in a store.
   - Experiments: Controlled studies where variables are manipulated to observe effects.
     Example: Clinical drug trials.

Data Collection Methods and Its Types

SLO 4.3.1 U

Types of Data Collection Methods

2. Secondary Data Collection: Data gathered from existing sources that were collected by others for different purposes.
   - Published Sources: Books, journals, reports, government publications.
     Example: Using census data for demographic studies.
   - Databases and Records: Organizational databases, online datasets, historical records.
     Example: Financial transaction records for market analysis.

Data Collection Methods and Its Types

SLO 4.3.1 U

Types of Data Collection Methods

3. Other Classification by Data Type

- Qualitative Data Collection: Focuses on non-numeric data capturing attitudes, opinions, or behaviors. Methods include interviews, focus groups, and observations.
- Quantitative Data Collection: Involves numeric data measurable and analyzable statistically. Methods include surveys with closed-ended questions, experiments, and existing numerical datasets.

SLO 4.3.2 An

analyse the role of data science in addressing real-world problems, supported by practical examples.

SLO 4.3.2 An

Role of Data Science in Solving Real-World Problems

- Data science combines statistical methods, machine learning, and domain knowledge to extract insights from data, enabling informed decision-making and problem-solving across various fields.

SLO 4.3.2 An

Role of Data Science in Solving Real-World Problems

1. **Improving Healthcare Outcomes**: Data science enables predictive modeling for early disease detection, personalized treatments, and resource optimization.
   Example: Using machine learning algorithms to predict patient readmission risks, allowing hospitals to provide targeted care and reduce costs.
2. **Enhancing Business Decisions**: Companies analyze customer behavior, market trends, and operational data to optimize marketing, inventory, and pricing strategies.
   Example: E-commerce platforms use recommendation systems powered by data science to increase sales by suggesting relevant products to customers.
3. **Optimizing Supply Chains**: Data-driven forecasting and logistics models improve inventory management and delivery efficiency.
   Example: Amazon employs data science to predict demand and optimize warehouse stocking and shipping routes.

SLO 4.3.2 An

Role of Data Science in Solving Real-World Problems

4. **Fraud Detection and Security**: Unsupervised learning models detect anomalies indicating fraudulent transactions or cybersecurity breaches.
   Example: Banks use data science to monitor transaction patterns and flag suspicious activity in real time.
5. **Environmental Monitoring and Climate Change**: Data science models analyze environmental data to predict weather patterns, track pollution, and manage natural resources.
   Example: Predictive models forecast hurricane paths, helping authorities plan evacuations and reduce damage.
6. **Social Good and Public Policy**: Governments and NGOs use data science for public health surveillance, crime prediction, and resource allocation.
   Example: Analyzing social media and mobility data to track disease outbreaks and inform vaccination campaigns.

ANY
Questions?

Thank You!