NIRMA UNIVERSITY

Institute:	Institute of Technology, School of Technology
Name of Programme:	MTech (Data Science)
Course Code:	6CS303CC25
Course Title:	Statistics for Data Science
Course Type:	Core
Year of Introduction:	2025-26

L	T	Practical Component				C
		LPW	PW	W	S	
3	0	2	-	-	-	4

Course Learning Outcomes (CLO):

At the end of the course, the students will be able to:

1. apply statistical concepts such as descriptive statistics, probability distributions, and inferential techniques to analyze and summarize data (BL3)
2. evaluate hypotheses and perform statistical tests to make data-driven decisions (BL5)
3. interpret and visualize statistical results effectively for decision-making in data science applications (BL4)
4. formulate and solve optimization problems using Linear Programming. (BL3)

Unit	Contents	Teaching Hours (Total 45)
Unit-I	**Descriptive Statistics & Data Visualization:** Measures of central tendency, Measures of dispersion, Data visualization techniques (Histograms, Boxplots, Scatter plots, Bar charts), Exploratory Data Analysis (EDA)	06
Unit-II	**Probability Theory & Distributions:** Basics of probability (Rules, Conditional Probability, Bayes Theorem), Discrete and continuous probability distributions, Common distributions: Binomial, Poisson, Normal, Exponential, Central Limit Theorem (CLT)	09
Unit-III	**Statistical Inference & Hypothesis Testing:** Point Estimation, Sampling methods and sampling distribution, Confidence intervals, Correlation and covariance, Hypothesis testing (t-test, z-test, ANOVA, Chi-square test), Type I & II Errors, p-value interpretation, Markov chains, Hidden Markov Models	15

Unit-IV	**Optimization Techniques:** Introduction to Optimization, Role of optimization in data science and machine learning, Types of optimization problems: Unconstrained vs. Constrained, Linear & Nonlinear Optimization: Linear Programming (LP) and Simplex Method, Quadratic and Nonlinear Programming, least square optimization, Gradient-Based Optimization, Convex Optimization.	15

Self-Study:

The self-study contents will be declared at the commencement of the semester. Around 10% of the questions will be asked from self-study content.

Suggested Readings/ References:

1. Montgomery, D. C., & Runger, G. C., Applied Statistics and Probability for Engineers. Wiley.
2. Rice, J. A., Mathematical Statistics and Data Analysis. Cengage Learning.
3. Wasserman, L., All of Statistics: A Concise Course in Statistical Inference. Springer.
4. Hastie, T., Tibshirani, R., & Friedman, J., The Elements of Statistical Learning. Springer.

Suggested List of Experiments:

Sr. No.	Name of Experiments/Exercises	Hours
1	Given a dataset (e.g., student exam scores), compute the mean, median, mode, variance, and standard deviation. Analyze how these measures change with different datasets (skewed vs. normal distributions) and interpret the results	02
2	Data Visualization Techniques (Histograms, Boxplots, Scatter Plots, Bar Charts): Use a financial dataset (e.g., stock prices) to create a boxplot. Identify outliers and extreme values, and interpret their significance. Discuss what might cause these outliers and their impact on data analysis	02
3	Exploratory Data Analysis (EDA) on a Real-World Dataset: Perform exploratory analysis on Titanic or any other dataset	02
4	Analyze a dataset with multiple variables (e.g., sales vs. advertising budget, temperature vs. humidity). Compute the correlation matrix and create a heatmap to visualize relationships. Explain how these correlations can impact decision-making	02
5	Simulating Discrete & Continuous Probability Distributions: Simulate 1,000 trials of flipping a fair coin. Use the binomial distribution to compute the expected number of heads and tails. Compare the simulated and theoretical results, and discuss the concept of expected value	02
6	Hypothesis Testing (t-test, z-test, Chi-square test) on relevant use cases	04
7	Implement Viterbi algorithm for addressing the decoding problem in hidden Markov models	04

8	Given a sequence of weather observations (e.g., sunny, cloudy, rainy), and a set of possible hidden states (e.g., "high pressure system", "low pressure system"), use a Hidden Markov Model to predict the next weather state	04
9	Conduct an ANOVA test to compare the means of three or more groups and determine if at least one of the group means is significantly different from the others	04
10	Gradient Descent optimization for machine learning models.	04