Course Code				23STMAP236			
Title of the Course				Statistical Data Analysis using R			
Offered to:				BSc.Honour – Artificial Intelligence			
L	0	T	0	P	2	C	1
Year of Introduction:	2024-25			Semester:		3	
Course Category:	MAJOR			Course Relates to:	Local, Regional, National, Global		
Year of Revison	NA			Percentage:	NA		
Type of the Course:				SKILL DEVELOMENT			
Crosscutting Issues of the Course :				NA			
Pre-requisites, if any				Basic Mathematics			

Course Description:

This lab course introduces students to the application of R programming in data analysis, focusing on visualization, descriptive statistics, and probability distribution modeling. Through a series of hands-on labs, students will learn to create visual representations of data, compute key descriptive statistics, and model both discrete and continuous probability distributions such as Binomial, Poisson, Uniform, Exponential, and Normal distributions. Additionally, students will apply linear regression and correlation analysis techniques to explore relationships between variables, enhancing their data interpretation and decision-making skills.

Course Objectives:

S. No	**COURSE OBJECTIVES**
1	**Apply R programming for creating diagrammatic representations** such as bar plots, histograms, and box plots for effective data visualization and interpretation.
2	**Compute and interpret descriptive statistics** (mean, median, variance, skewness, kurtosis) using R, summarizing data distributions and understanding their characteristics.
3	**Simulate and analyze discrete probability distributions** (Binomial and Poisson) using R, comparing empirical results with theoretical models.
4	**Model and evaluate continuous probability distributions** (Uniform, Exponential, and Normal) using R, applying statistical methods to assess and compare empirical data with theoretical expectations.
5	**Apply correlation and regression analysis techniques** using R to explore relationships between variables, and predict outcomes based on statistical models.

Course Outcomes

At the end of the course, the student will be able to…

NO	COURSE OUTCOME	BTL	PO	PSO
CO1	apply R programming to create various diagrammatic representations such as bar plots, histograms, and box plots, enabling effective visualization and interpretation of datasets.	K3	2	2
CO2	demonstrate the ability to compute and interpret key descriptive statistics (mean, median, variance, skewness, and kurtosis) using R, allowing them to summarize data distributions and understand their characteristics.	K3	2	2
CO3	simulate and analyze binomial and Poisson distributions using R, and compare theoretical and empirical results by calculating probabilities, means, and variances.	K3	2	2
CO4	develop skills in modeling continuous distributions (Uniform, Exponential, and Normal) using R, applying statistical methods to calculate and compare empirical data with theoretical distributions.	K3	2	2
CO5	apply linear regression and correlation techniques in R to analyze relationships between variables and interpret the results to understand underlying patterns and predict outcomes.	K3	2	2

For BTL: K1: Remember; K2: Understand; K3: Apply; K4: Analyze; K5: Evaluate; K6: Create

CO-PO-PSO MATRIX									
CO NO	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PSO1	PSO2
CO1		3							3
CO2		3							3
CO3		3							3
CO4		3							3
CO5		3							3

Use the codes 3, 2, 1 for High, Moderate and Low correlation Between CO-PO-PSO respectively

Course Structure

I. Data Visualization & Diagrammatic representation (K3) 4Hrs

Lab1: Diagrammatic representation using R

Experiment:

1. Creating a dataset for products and their sales.
2. create a dataset of 100 exam scores for students, ranging between 0 and 100.
3. create a simple dataset of market shares for five companies.

Tasks:

1. The basic task is to visualize the frequency or value for different categories. In R, the bar plot () function is used for this.
2. The basic task is to create a pie chart that shows proportions for each category.
3. Create a Basic Histogram Use the hist () function to create a histogram.
4. Boxplots are often used to compare the distribution of a variable across different groups.
5. Boxplots are excellent for identifying outliers.

II. Descriptive Statistics (K3) 4Hrs

Lab2: Descriptive statistics using R

Experiment:

1.Analyzing the Heights of a Group of People.

2. Analysing Daily Temperature Fluctuations.

3. Analysing Monthly Sales Data

Tasks:

1. Calculate the mean, median and mode of the data set.

2. Calculate the variance, standard deviation, etc.

3. Calculate the skewness and kurtosis and observe the nature of the distribution.

III. Binomial Distribution(K3) 2Hrs

Lab 3: Modelling Coin Tosses with a Binomial Distribution

Dataset: Simulated or real-world data on coin tosses.

Experiment:

1. Simulate a series of coin tosses using a random number generator.
2. Count the number of heads and tails.
3. Calculate the probability of success (e.g., heads) based on the simulation results.

Tasks:

1. Calculate the mean and variance of the binomial distribution.
2. Compare the simulated results with the theoretical values for a binomial distribution.
3. Explore the effect of changing the number of trials and the probability of success on the distribution.

Lab 4: Fitting of Binomial distribution (K3)

Dataset: Real world data set

Tasks:

1. Calculate the mean and variance of the binomial distribution and also calculate Probability of success.
2. Compare the expected results using recurrence relation for probabilities with the theoretical values for a binomial distribution.

IV. Poisson Distribution (K3) 2Hrs

Lab 5: Simulating and analysing a Poisson Distribution

Dataset: Simulated data generated in R.

Experiment:

1. Generate a random sample from a Poisson distribution using the **rpois** function in R.

2. Calculate the mean, variance, and standard deviation of the sample.

Tasks:

1. Compare the calculated statistics with the theoretical values for a Poisson distribution.
2. Create a histogram to visualize the distribution of the simulated data.
3. Explore the effect of changing the rate parameter (lambda) on the distribution.

Lab 6: Fitting of Poisson distribution

Dataset: Real world data set

Tasks:

1. Calculate the mean and variance of the Poisson distribution and also calculate lambda value.
2. Compare the expected results using recurrence relation for probabilities with the theoretical values for a Poisson distribution.

V. Continuous Uniform distribution (K3) 2Hrs

Lab 7: Understanding and Generating Continuous Uniform Distributions in R

Title: Exploring the Continuous Uniform Distribution in R

Dataset: Simulated data generated using R's runif() function.

Experiment: Investigate the properties of the continuous uniform distribution and its applications in R.

Tasks:

1. **Generate uniform random numbers:** Create a vector of 1000 random numbers from a uniform distribution with specified minimum and maximum values.
2. **Visualize the distribution:** Plot a histogram and density plot of the generated random numbers to visualize the uniform distribution.
3. **Calculate statistical properties:** Compute the mean, variance, and standard deviation of the generated random numbers.
4. **Compare with theoretical values:** Verify that the calculated statistical properties match the theoretical values for a uniform distribution.
5. **Applications:** Explore real-world applications of the uniform distribution, such as random sampling, Monte Carlo simulations, or hypothesis testing.

Lab 8: Simulating and Analyzing Uniformly Distributed Data

Title: Simulating and Analyzing Uniformly Distributed Data in R

Dataset: Simulated data generated using R's runif() function.

Experiment: Simulate and analyze data from a continuous uniform distribution to understand its characteristics and applications.

Tasks:

1. **Generate random samples:** Create multiple samples of varying sizes from a uniform distribution with specified parameters.
2. **Calculate statistics:** Compute the mean, variance, and standard deviation for each sample.
3. **Visualize the distribution:** Plot histograms and density plots for each sample to observe the distribution's shape and variability.

VI Exponential Distribution 2Hr

Lab 9: Understanding and Generating Exponential Distributions in R

Title: Exploring the Exponential Distribution in R

Dataset: Simulated data generated using R's rexp() function.

Experiment: Investigate the properties of the exponential distribution and its applications in R.

Tasks:

1. **Generate exponential random numbers:** Create a vector of 1000 random numbers from an exponential distribution with a specified rate parameter (λ).
2. **Visualize the distribution:** Plot a histogram and density plot of the generated random numbers to visualize the exponential distribution.
3. **Calculate statistical properties:** Compute the mean, variance, and standard deviation of the generated random numbers.
4. **Compare with theoretical values:** Verify that the calculated statistical properties match the theoretical values for an exponential distribution.
5. **Applications:** Explore real-world applications of the exponential distribution, such as modeling waiting times, reliability analysis, or queuing systems.

Lab 10: Simulating and Analyzing Exponentially Distributed Data

Title: Simulating and Analyzing Exponentially Distributed Data in R

Dataset: Simulated data generated using R's rexp() function.

Experiment: Simulate and analyze data from an exponential distribution to understand its characteristics and applications.

Tasks:

1. **Generate random samples:** Create multiple samples of varying sizes from an exponential distribution with specified rate parameters.
2. **Calculate statistics:** Compute the mean, variance, and standard deviation for each sample.

3. **Visualize the distribution:** Plot histograms and density plots for each sample to observe the distribution's shape and variability.
4. **Hypothesis testing:** Conduct hypothesis tests to determine if a given sample comes from an exponential distribution.

VII Normal Distribution 4Hrs

Lab 11: Understanding and Generating Normal Distributions in R

Title: Exploring the Normal Distribution in R

Dataset: Simulated data generated using R's rnorm() function.

Experiment: Investigate the properties of the normal distribution and its applications in R.

Tasks:

1. **Generate normal random numbers:** Create a vector of 1000 random numbers from a normal distribution with specified mean and standard deviation.
2. **Visualize the distribution:** Plot a histogram and density plot of the generated random numbers to visualize the normal distribution.
3. **Calculate statistical properties:** Compute the mean, variance, and standard deviation of the generated random numbers.
4. **Compare with theoretical values:** Verify that the calculated statistical properties match the theoretical values for a normal distribution.
5. **Applications:** Explore real-world applications of the normal distribution, such as modeling measurement errors, financial returns, or population characteristics.

Lab 12: Simulating and Analyzing Normally Distributed Data

Title: Simulating and Analyzing Normally Distributed Data in R

Dataset: Simulated data generated using R's rnorm() function.

Experiment: Simulate and analyze data from a normal distribution to understand its characteristics and applications.

Tasks:

1. **Generate random samples:** Create multiple samples of varying sizes from a normal distribution with specified mean and standard deviation.
2. **Calculate statistics:** Compute the mean, variance, and standard deviation for each sample.
3. **Visualize the distribution:** Plot histograms and density plots for each sample to observe the distribution's shape and variability.
4. **Hypothesis testing:** Conduct hypothesis tests to determine if a given sample comes from a normal distribution (e.g., Shapiro-Wilk test).

VIII Correlation coefficient **2Hr**

Title: Calculate the correlation for bivariate data in R

Dataset: Simulated data generated using R's cor() function.

Experiment: The basic example of calculating the Pearson correlation coefficient between two variables in the mtcars dataset.

Tasks:

1. Load the Data: Use the mtcars dataset in R.
2. Compute the Pearson Correlation Coefficient.
3. Compute the Spearman Correlation Coefficient.
4. Compute the Correlation Matrix.

Simple Linear Regression **2Hrs**

Title: Build a simple linear regression using R

Dataset: We will perform a simple linear regression to model the relationship between miles per gallon (mpg) and horsepower (hp) in the mtcars dataset.

Experiment: You will conduct a simple linear regression analysis to examine the relationship between two variables: miles per gallon (mpg) and horsepower (hp) from the mtcars dataset.

Tasks:

1. **Load the Dataset**, Load a built-in dataset, such as mtcars, and explore its structure.
2. **Explore the Data,** Create a scatter plot of mpg vs. wt, and overlay the regression line.
3. **Fit the Simple Linear Regression Model,** Fit a linear regression model to predict mpg (miles per gallon) using wt (weight of the car).
4. **Interpretation** of the Results.

IX Multiple Linear Regression **6Hrs**

Title: Build a multiple linear regression using R

Dataset: Perform a multiple linear regression to predict mpg using horsepower (hp), weight (wt), and cylinders (cyl) as predictors.

Experiment: The mtcars dataset to perform a multiple linear regression with miles per gallon (mpg) as the response variable and multiple predictors**.**

Tasks:

1. **Load the Dataset**, Load a built-in dataset, such as mtcars, and explore its structure.
2. **Explore the Data,** Create a scatter plot of mpg vs. wt, and overlay the regression line.

3. **Fit the Simple Linear Regression Model,** Fit a linear regression model to predict mpg (miles per gallon) using wt (weight of the car).

4. **Interpretation** of the Results.

Lab Manual:

1. Statistical Data Analysis using R programming –Sri K. Siva Naga Raju, 2024 First Edition, Department of Statistics, PBSCAS

References:

1. Zaven A. Karian, Edward J. Dudewicz, 2010, **Handbook of Fitting Statistical Distributions with R, 1st edition, Chapman and Hall/CRC.**
2. Vito Ricci, 2005, FITTING DISTRIBUTIONS WITH R, R-Cran Projects. https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf

***

SEE (LAB) Model Question Paper

23STMAP236: Statistical Data Analysis using R

Offered to B.Sc. Hons Artificial Intelligence

Max. Marks: 50 **Max. Time: 3Hrs**

Pass. Min: 20

(A)	**Evaluation Procedure**	**35 Marks**
I	Experiments (Exam & Execution)	**30 Marks**
II	Viva	**3 Marks**
III	Record	**2 Marks**
(B)	**CONTINUOUS ASSESMENT(Internal)**	**15 MARKS**

15 marks for the continuous assessment (Day to day work in the laboratory shall be evaluated for 15 marks by the concerned laboratory teacher based on the regularity/ record/viva). Laboratory teachers are mandated to ensure that every student completes 80%-90% of the lab assessments.

TOTAL: (A)+(B) = 50 MARKS