22DS2T1: ESSENTIALS OF STATISTICS FOR DATA SCIENCE USING R

Course Name	Essentials of Statistics for Data Science Using R	**L**	**T**	**P**	**C**	**CIA**	**SEE**	**TM**
Course Code	22DS2T1	4	0	0	4	30	70	100
Year of Introduction: 2021	**Year of Offering:** 2021	**Year of Revision**: 2022			**Percentage of Revision:** 10			
L-Lecture, T-Tutorial, **P**-Practical, **C**-Credits, **CIA**-Internal Marks, **SEE**-External Marks, **TM**-Total Marks								

Course Description and Purpose:

Essentials of Statistics for Data Science using R (22DS2T1) is a course that illustrates basic concepts of *R Programming*, *Bi-variate Analysis, Probability, Regressions, Time Series Analysis, Hypothesis Testing ,Analysis of ANOVA ,Connecting to R External Interfaces.*

Course Objectives:

This course will help enable the students o understand, learn and implement concepts of Statistics using R programming like *Bi-variate Analysis, Probability, Regressions, Time Series Analysis ,Hypothesis Testing ,Analysis of ANOVA ,Connecting to R External Interfaces.*

Course Objectives:

The learning objectives include:

- To understand basic concepts of *Statistics, R Programming and Bi-Variate Analysis*.
- To understand the concepts of *Probability, Random Variables and Probability Distribution and its Applications.*
- To understand and gain knowledge on *Regressions, Time Series of Analysis*
- To understand the concepts of *Hypothesis Testing and Analysis of ANOVA.*
- To understand how to import *Different Files* and *Connecting Databases to R*.

Course Outcomes:

After completing this course, the students should have developed a clear understanding of

CO1: Understand basic concepts of *Statistics, R Programming and Bi-Variate Analysis.*

CO2: Understand the concepts of *Probability, Random Variables and Probability Distribution and its Applications.*

CO3: Understand and gain knowledge on *Regressions, Time Series of Analysi.*

CO4: Understand the concepts of Hypothesis *Testing and Analysis of ANOVA.*

CO5: Understand how to *import Different Files* and *Connecting Databases to R.*

UNIT I (12 Hours)

Introduction to Statistics: Statistics Definition - Types of Statistical Methods - Data Collection (Definition , Sources of Data Collection, Methods of Data Collection) - Classification- Basic of Classification Types - Tabulation of Data (Meaning and Definition, Objectives, Types of Tables) - Exploratory Data Analysis (Types of Data Visualization).

Introduction to R Programming: Basic Data Types - Operations on Data Structures - Descriptive Statistics with R-Measures(Central Tendency and Measures of Dispersion of Variability).

Bi-variate Analysis using R: Correlation Meaning - Types of Correlation (Measures or Methods of Correlation, Scatter Diagram, Karl Pearson’s Coefficient of Correlation, Spearman’s Rank Correlation Coefficient) - Bivariate Analysis of Categorical Variables and numerical variables.

UNIT II (12 Hours)

Probability Using R: Various Definitions - Addition Theorem - Conditional Probability - Multiplication Theorem - Bayes' Theorem and its Applications - Random Variables: Definition, Discrete and Continuous Random Variables - Distribution Function and its Properties - Discrete Probability Distributions: Binomial, Poisson and Geometric - Continuous Probability Distributions - Uniform, Normal and Exponential Distributions - Properties and Applications.Applications of Probability using R.

UNIT III (12 Hours)

Regression: Introduction - Estimation the Method of Least Square - Regression Coefficients(Properties of Regression Coefficients, Coefficient of Simple Linear Determination) - Types of Regression Models (Simple Linear Regression , Multiple Linear Regression, Logistic Regression) - Assumptions of Regression Models, Applications and its implementation using R Programming

Time Series Analysis using R: Meaning of Time Series - Components Of Time Series - Time Series Decomposition Models (Multiplicative Model and Additive Model) - Forecasting Methods (Simple Moving Averages and Weighted Moving Averages).

Note: Proofs and derivations of statements are excluded.

UNIT IV (12 Hours)

Testing of Hypothesis Using R: Definition of Hypothesis - Steps in Testing of Hypothesis - Types of Hypothesis Testing - Hypothesis Testing of Means and Proportions - Testing for Differences between Means and Proportions.

Non Parametric Tests: The MannWhitney U Test - Kruskal Wallis Test - Wilcoxon Signed Rank Test and Chi Square Test.

Analysis of Variance Using R: One way ANOVA - Two way ANOVA - Multivariate Analysis of Variance (MANOVA).

UNIT V (12 Hours)

Connecting R to External Interfaces: CSV Files (Reading From a CSV File, Writing to a CSV File) - Microsoft Excel (Reading from XLSX File, Writing to XLSX File) - Databases (Connecting R to MYSQL (Creating Tables, Inserting Rows, Updating Rows, Deleting Rows, Querying Rows, Querying Tables, Dropping Tables)) - XML Files (Reading From XML Files, JSON Files, Reading From JSON Files), Binary Files (Writing to Binary Files, Reading From Binary Files).

Reference Text Books:

1. Sharma, J. K., Business Statistics (UNIT-I,UNIT-III), New Delhi: Pearson Education, 2013
2. Anderson,D.,Sweeney,D.,Williams,T., Camm, J., & Cochran, J., Statistics for Business and Economics, Cengage Learning, 2013, New Delhi
3. Dr. Rob Kabacoff, R in Action: Data Analysis and Graphics with R (UNIT-IV), Manning Publications CO, Edition 2011.
4. Dr.Jeeva Jose, A Beginners Guide for Data Analysis Using R Programming. (UNIT-II, UNIT-V, UNIT-III), Khanna Book Publishing Co.(P) Ltd, Edition 2019.
5. Michael J. Crawley, John Wiley & Sons, Statistics: An Introduction using R, Weily, 2015.
6. Aczel,A.D.& Sounderpandian, J, Complete Business Statistics, Tata McGraw Hill, 2011, New Delhi.
7. Davis, G., & Pecar, B., Business Statistics using Excel, New Delhi: Oxford University Press, 2014.

22DS2T1

P.B.SIDDHARTHA COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS),
VIJAYAWADA-520010

(An Autonomous College in the Jurisdiction of Krishna University, A.P., India.)

M.Sc.,(Computational Data Science) DEGREE EXAMINATIONS
SECOND SEMESTER
ESSENTIALS OF STATISTICS FOR DATA SCIENCE USING R
SYLLABUS W.E.F 2022-2023

Time 3 Hours **Max.Marks: 70**

Answer all questions. All question carry equal marks. **5 × 4 Marks =20 Marks**

1.(a) Explain types of *Statistical Methods*.(CO1,L2)

(OR)

(b) Explain *Types of Correlation* with examples. (CO1,L2)

2.(a) Explain *Distribution Function* and its Properties. (CO2,L2)

(OR)

(b) Explain *Applications of Probability* using R. (CO2,L2)

3. (a) How we can determine the Coefficients of *Simple Linear Regression*? (CO3,L1)

(OR)

(b) What are the components of *Time Series*. (CO3,L1)

4. (a) What are the steps involved in *Hypothesis Testing*. (CO4,L1)

(OR)

(b) What is meant by *Two Way ANOVA*? Give one example using R .(CO4,L1)

5. (a) How can you create table and insert rows in table with the help of MYSQL using R. (CO5,L1)

(OR)

(b) How do you import *XML Files* using R with example? (CO5,L1)

Answer the following **5 × 10M = 50Marks**

1.(a) What is *Descriptive Statistic*? Explain about *Measures of Central Tendency* and *Dispersion of Variability* using R. (CO1,L1) 10 Marks

(or)

(b) What is *Correlation*? Explain *Karl Pearson's Coefficient* and *Spearman's Rank Correlation Coefficient* using R. (CO1,L1) 5 Marks

(c) What is *Bi-variate Analysis*? How we can implement using categorical and numerical data using R? (CO1,L1) 5 Marks

2. (a) Explain *Addition Theorem of Probability* using an example. (CO2,L2) 5 Marks

(b) Illustrate *Conditional Probability*? Explain *Baye's Theorem* without Proof. (CO2,L2) 5Marks

(or)

(c) Explain the assumption of *Poisson Distribution* and give its *Probability Distribution Function* using R with example (CO2,L5) 5 Marks

(b) Explain the properties *of Normal Distribution* and give its *Probability Distribution Function using R*. (CO2,L5) 5Marks

3.(a) Construct different *Regression Models* using R. (CO3,L3) 10 Marks

(or)

(c) Apply *Simple Moving Averages* and *Weighted Moving Averages* using R. (CO3,L3) 10 Marks

4. (a) List any two approaches used in *Non Parametric Testing.* (CO4,L4) 10 Marks

(or)

(b) Analyze *Hypothesis Testing of Means and Proportions* and its differences with examples using R. (CO4,L4) 10 Marks

5.(a) Develop database connection in R using MYSQL commands? Give one example. (CO5,L6) 10 Marks

(or)

(b) Discuss about JSON files and binary files in R with examples? (CO5,L6) 10 Marks