Generic Elective (GE) Courses

Generic Elective (GE) Course 1a: Statistical Computing Using R

Course Title & Code	Credits	Credit Distribution of the Course			Eligibility Criteria	Prerequisite of the course (if any)
		Lecture (45 Hours)	Tutorial (00 Hours)	Practical (30 Hours)		
GE 1a: Statistical Computing Using R	4	3	0	1	NIL	NIL

Course Objectives:

- To learn the principles and methods of data analysis.
- To provide a basic understanding of methods of analysing data from different fields.
- To perform data analysis using R software.

Course Learning Outcomes: After successful completion of this course, the students will be able to:

- Carry out data analysis using R software.
- Effectively visualize and summarize the data.
- Interpret the results of statistical analysis.

Unit I (11 Hours)

Introduction to R: Installing R, R console, Script file, Workspace, Getting help, R packages, Installing and loading packages. R data structures: vectors, matrices, array, data frames, factors, lists. Creating datasets in R, Importing and exporting dataset, annotating datasets. Graphs: Creating and saving graphs, customizing symbols, lines, colors and axes, combining multiple graphs into one, bar plots, boxplot and dot plots, pie chart, stem and leaf display, histogram and kernel density plots. Data management: Manipulating dates and missing values, understanding data type conversion, creating and recoding variables, sorting, merging and sub-setting data sets. Mathematical and statistical functions, character functions, looping and conditional statements, user defined functions.

Unit II (11 Hours)

Basic statistics: Descriptive statistics, frequency and contingency tables, outlier detection, testing of normality, basics of statistical inference in order to understand hypothesis testing, p-value and confidence intervals. Parametric tests: Tests for population mean and variance for two or more populations, tests for independence and measures of association, sample size determination for common statistical methods using pwr package. Nonparametric tests.

Unit III (11 Hours)

Correlation: Correlations between quantitative variables and their associated significance tests. Regression Analysis: Fitting simple and multiple regression model forward, backward and stepwise regression, polynomial regression, regression diagnostics to assess the statistical assumptions, methods for modifying the data to meet these assumptions more closely, selecting a final regression model from many competing models. ANOVA: Fitting and interpreting ANOVA type models, evaluating model assumptions, basic experimental designs: CRD, RBD, LSD and factorial experimental designs.

Unit IV (12 Hours)

Time series Analysis: Creating and manipulating a time series, Components of a time series, auto-correlation and partial correlation function, estimating and eliminating the deterministic components of a time series. Developing Predictive Models: Forecasting using exponential models, predictive accuracy measures for time-series forecast, testing for stationarity, Forecasting using ARMA and ARIMA models. EM algorithm: Applications to missing and incomplete data problems, mixture models.

Essential Readings:

1. Davies, T.M. (2016). *The Book of R: A First Course in Programming and Statistics*, No Starch Press.
2. Field, A., Miles, J. and Field, Z. (2012). *Discovering Statistics using R*, Sage.

Suggested Readings:

1. Crawley, M.J. (2013). *The R Book*, John Wiley & Sons.
2. Kabacoff, R.I. (2015). *R in Action: Data Analysis and Graphics in R*, Manning Publications.
3. Rizzo, M.L. (2019). *Statistical Computing with R*, Chapman & Hall/CRC Press.

List of Practicals:

1. Problems based on creating vectors and mathematical operations.
2. Problems based on sequences, replications, sorting and lengths.
3. To perform matrix operations, importing and exporting datasets.
4. Basic plotting of R graphical functionality.
5. Basic statistics and testing of hypothesis.
6. Parametric and non-parametric tests.
7. Problems based on sample size determination.
8. Correlation and regression analysis using quantitative variables.
9. Analysis of variance and basic design of experiments (CRD, RBD, LSD and factorial designs).
10. To plot a time series function, autocorrelation function and correlogram.
11. Problems based on ARMA and ARIMA models.
12. Problems based on EM algorithms for missing and incomplete data problems.