GUJARAT TECHNOLOGICAL UNIVERSITY

Syllabus for Master of Computer Applications, 2nd Semester
Subject Name: Statistical Methods
Subject Code: 629409

With effective from academic year 2020-21

1. **Learning Objectives:**
   To understand and apply various concepts, techniques and methods used in Descriptive Statistics and Inferential Statistics. The knowledge and skills gained will equip students in carrying out preliminary Data Analytics tasks, and to prepare foundation to understand and apply the statistical techniques in various fields such as Total Quality Management, Simulation, Game Theory, Operations Research, etc. in addition to Computer Science topics such as Machine Learning, Cryptography, Artificial Intelligence, Operating Systems, Data Structures and Algorithms, etc.

2. **Prerequisites**: Preliminary mathematical concepts

3. **Contents**:

Unit	Course Content	Weightage Percentage
Unit I	**Introduction to Statistics and Descriptive Statistics** Introduction, Broad areas (classification) of Statistics; **Describing Data Visually:** Frequency Distributions and Histograms; Pie Charts; Bar Charts: Pareto Chart, Scatter Plots (Degree of Association); Line Charts; **Descriptive Statistics:** Central Tendency; Mean and its Characteristics, Median and its Characteristics, Quartiles and Percentiles, Mode; **Dispersion:** Range, Mean Absolute Deviation, Interquartile Range (IQR); Variance, Standard Deviation and its Characteristics, Coefficient of Variation; **Standardized Data:** including Chebyshev's Theorem, Outliers; **Box Plots:** including Fences and Unusual Data Values; **Grouped Data:** Nature, Mean and Standard Deviation, Accuracy Issues; **Skewness:** Coefficient of Skewness; **Kurtosis**: Leptokurtic, Platykurtic, Mesokurtic; **Measures of Association:** Covariance, Correlation, Coefficient of Correlation; Correlation and Causation	**18%**

Unit II	**Probability and Probability Distributions** **Introduction:** Common Framework: Experiment, Event, Elementary Events, Sample Space; Definition of Probability; Marginal Probability; Probability of Union of Events (Addition Laws), Probability Matrix; Probability of Complement of a Union; Probability of Joint Events (General Laws of Multiplication); Conditional Probability; Mutually Exclusive Events, Independent Events; Revision of Probability Values: Bayes' Rule **Discrete Probability Distributions:** Introduction, Binomial Distribution, Poisson Distribution, Applications; **Continuous Probability Distributions:** Introduction, Normal Distribution, Exponential Distribution, Applications;	**24%**
Unit III	**Sampling, Sampling Distributions and Estimation** **Types of Sampling:** Random, Nonrandom; Sampling Distribution of x-bar; Central Limit Theorem; z Formula for Sample Mean; Standard Error of Mean; Sampling from a Finite Population; Sampling Distribution of a Proportion, Standard Error of Proportion **Estimation for Single Population:** Estimating the Population Mean using z Statistic (σ Known); Estimating the Population Mean using the z Statistic when the Sample Size is Small; Estimating the Population Mean using t Statistic (σ Unknown); Estimating the Population Proportion; Estimating the Population Variance; Estimating Sample Size	**24%**
Unit IV	**One Sample Hypothesis Tests** Introduction; Null Hypothesis, Alternate Hypothesis; Type I & Type II Errors, Testing Hypotheses about Population Mean using z Statistic (σ Known); Using Critical Value Method to test Hypotheses, Examples; Testing Hypotheses about Population Mean using t Statistic (σ Unknown); Testing Hypotheses about Proportion; Testing Hypotheses about Variance	**18%**
Unit V	**Regression** Introduction, Simple Regression Analysis, Least Square Analysis to Determine the Equation of Regression Line; Residual Analysis, Using Residual to Test the Assumptions of the Regression Model; Standard Error of the Estimate; Coefficient of Determination; Hypothesis Testing for the Slope of the Regression Model; Testing the Overall Model; Using Regression to Develop a Forecasting Trend Line	**16%**

Optional (but Recommended) Topics:

- Overview of other Discrete and Continuous Probability Distributions
- Overview of Statistical Inferences about Two Populations; Analysis of Variance
- Overview of Multiple Regression Model; Mathematical Transformation of Nonlinear Models

4. Text Book:

1) Ken Black, "Business Statistics for Contemporary Decision Making", Wiley Student Edition, 2010

5. Reference Books:

1) David P. Doane, Lori E. Seward, "Applied Statistics in Business and Economics" Tata McGraw-Hill, 2010
2) Anderson, Sweeney, Williams, "Statistics for business and economics", 9th edition, 3**)** Thompson Publication
4) Bharat Jhunjhunwala, "Business Statistics", first edition, S Chand, 2008
5) Richard Levin, David Rubin, "Statistics for Management", 7th edition, PHI
6) Nabendu Pal, Sahadeb Sarkar, "Statistics-Concepts and Applications", 2nd edition, PHI
7) J. Susan Milton & Jesse Arnold, "Introduction to Probability & Statistics: Principles & Applications for Engineering & Computing Sciences", McGraw-Hill Education
8) S P Gupta, "Statistical Methods", 30th edition, S Chand

6. Chapter wise coverage from the Text Books:

Unit#	Chapter #
I	Chapter 1,2,3
II	Chapter 4,5,6
III	Chapter 7,8
IV	Chapter 9- (9.1-9.5) ,Chapter 10-10.1 , Chapter 11-11.1
V	Chapter 12 (12.2-12.7,12.9), Chapter :13 -13.1, Chapter 14-14.1

7. Accomplishment of the student after completing the course:
Students will be able to apply various concepts, techniques and methods used in Descriptive Statistics and Inferential Statistics in carrying out preliminary Data Analytics tasks. They will also be able to apply the statistical techniques in various fields such as Total Quality Management, Simulation, Game Theory, Operations Research, etc. in addition to Computer Science topics such as Machine Learning, Cryptography, Artificial Intelligence, Operating Systems, Data Structures and Algorithms, etc.

Practical List

Objectives: To implement statistical concepts using a standard tool, such as R. Such implementation is aimed at improved understanding and visualization of theoretical concepts. It is also aimed at laying a foundation for Data Analytics and Data Science.

Prerequisites: Logical Thinking and Basic Statistical Concepts

Advice (Note) to Teachers:
The list of exercises given below is an indicative list.

Note: R has many datasets. Get the available datasets through command data (). Use R commands related to Statistics for several datasets for a good practice.

Some exercises have been labelled as "**Mandatory**" while other exercises have been marked as "**Desirable**". It is expected that all the students will do **Mandatory** exercises while bright students will additionally do **Desirable** exercises as well.

List of Computer Lab Exercises

1. Introduction and a quick tour to R and R Studio (to be done in Lab) [09 Hours]

- **(a)** Basic data structures and constructs
- **(b)** Available R Datasets, such as mtcars, faithful, etc.
- **(c)** Null, NA, Missing Values
- **(d)** Basic Packages related to Statistics: e. g. stats, stats4, graphics, grDevices, modeest, agricolae, etc.

2. Descriptive Statistics [09 Hours]

- **(a)** Compute Mean, Median, Quartiles, Percentile (use quantile () function), Variance, Standard Deviation, IQR, Minimum & Maximum Values, Summary Statistics &interpretation(**Mandatory**)
- **(b)** Histogram, Scatter Plot, Box Plot, Density Plot of R data sets and interpretation (**Mandatory**)
- **(c)** Generate Frequency Distribution of data as a data frame(**Mandatory**)
- **(d)** Compute Correlation Coefficient and Covariance (**Mandatory**)

3. Probability and Probability Distributions [09 Hours]

- **(a)** Use pnorm(), pbinom(), ppois(), pexp() functions to compute probabilities based on a specific Probability distribution. (**Mandatory**)
- **(b)** Use dnorm(), dbinom(), dpois(), dexp() functions to compute probability density functions (**Mandatory**)
- **(c)** Use qnorm(), qbinom(), qpois, qexp() functions to get x value corresponding to given probability value (**Mandatory**)
- **(d)** Use different parameter values in 3 (a), and 3 (b) to observe the impact of different parameter values and prepare a note on that.(**Mandatory**)
- **(e)** Plot above results and interpret (**Desirable**)
- **(f)** Statistical test for normality using shapiro.test() function (**Desirable**)

4. Sampling, Sampling Distribution, Hypothesis Testing [12 Hours]

(a) Random sampling with or without replacement using sample () function (**Mandatory**)

(b) Generate n random samples (take n = 10, 50, 100, 200, 500, 1000 as an example), create a vector of Sample Means. Draw the Density Plot of Sample Means to visualize Central Limit Theorem (**Mandatory**)

(c) Take a sample and carry out Hypothesis Testing for the following cases:

1. Std. Deviation known, Large Sample Size, Sample from Non-Normal Population
2. Std. Deviation known, Small Sample Size, Sample from Normal Population
3. Std. Deviation known, Small Sample Size, Sample from non-Normal Population
4. Std. Deviation not known
5. Hypothesis Test for Variance (Chi-square Test) (**Desirable**)

5. Regression and Linear Modeling [06 Hours]

(a) Linear regression: One Independent Variable using lm () function; Interpret the output of Model Analysis, Compute Correlation Coefficient, Interpret results (**Mandatory**)

(b) Linear regression: Multiple Independent Variables using lm () function; Interpret the output of Model Analysis (**Mandatory**)

6. Using Tree data set (Inbuilt data set of R studio) create

(a) Histogram with proper label
(b) Scatterplot
(c) Boxplot
(d) Dot Plot
(e) Density Plot

7. Draw the different charts for Google Play Store data set (use following link)

https://docs.google.com/spreadsheets/d/12Uy4zsoUR44GfgVRtTIXRPr2DjwbW_-1rQN3lMextkg/edit?usp=sharing

(a) Rate wise Application
(b) Application popularity
(c) Number of users installing specific application
(d) Paid or Free Application
(e) Price wise Application
(f) Categorized Application (Entertainment, game, Education, Insurance etc.)
(g) Age group wise Application (Children, Teen, Mature, adult)

8. Perform Linear regression using employee dataset (use following link for dataset)
https://docs.google.com/spreadsheets/d/1r73YhZYxO1GiXZbiu1BREgJLlSyYhlv0y4jVLN2Niz4/edit?usp=sharing

Reference Books:

1. Pierre-Andre Cornillon, Arnaud Guyader, Francois Husson, Nicolas Jegou, Julie Josse, Maela Kloareg, Eric Matzner-Lober, Laurent Rouvière, "R for Statistics", CRC Press.
2. Dr. Mark Gardener, "Beginning R: The Statistical Programming Language", Wiley.
3. Paul Teetor, "R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics", O'Reilly Cookbooks.

Reference Websites:

1. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
2. https://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf
3. https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/reading-questions-r-intro
4. https://www.datacamp.com/introduction-to-statistics
5. http://tut-dl.com/item/lynda-r-statistics-essential-training
6. https://www.analyticsvidhya.com › Machine Learning
7. https://www.coursera.org/learn/r-programming
8. https://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-mathematicsdata-science/

Accomplishment of the student after completing the course:

1. Students will be able to carry out preliminary data analysis with results displayed graphically, and study the characteristics of standard probability distributions with their plots.
2. Students will also be able to demonstrate the inductive proof of Central Limit Theorem and go through linear regression (model) with fitness test of model.