# Statistical and Mathematical Methods for Data Analysis

## Rationale:

Data analysis and deducing the inferences for decision making are commonly encountered problems in experimental work both in the real field and simulation on computers. This course's objective is to impart to the students working knowledge of frequently required statistical methods with a particular emphasis on data science.

## Aims and Objectives:

- Students will get in-depth knowledge of how to do statistical analysis and apply them in real-life scenarios.
- After the completion of this course, students should be able to apply probability and statistics, especially in data science.
- This course helps students in doing statistical analysis in their areas of expertise.
- Students should be able to apply the learned concepts in Python.
- Students will learn some important Python libraries like NumPy, Matplotlib, Pandas, Scipy.
- It is expected from a student to write/submit their min project report in a conference/journal paper format.

## Catalog Description:

This is an advanced course on Statistical and Mathematical Methods for Data Analysis. The course will focus on basic concepts of sets and probability, random variables and probability distributions, fundamental sampling distributions and data descriptions, one- and two-sample estimation problems, (one- and two-sample tests of hypotheses), simple linear regression and correlation, multiple linear regression and certain nonlinear regression models estimating the coefficients, time series, three-sigma rule, the law of large numbers, simulation and application of all learned concepts in Python.

## Pre-requisite(s):

None

Course Outline:

## I. Basic concepts of Sets and Probability

1. Introduction, sets, relations, Venn Diagrams, complement, union, intersection, DE Morgan's law, cartesian products, power sets, cardinality, infinity, countability.

2. Counting, permutations, combinations, binomial coefficients and theorem. Probability, events, probability of unions, intersections. Uniform probability spaces.

3. Sampling with and without replacement, non-uniform probability spaces, axioms of probability.

4. Conditional probability, multiplication rule, probability tree diagrams, law of total probability, dependence and independence, posterior probability, law of total expectation, Bayes' theorem, probability density function (PDF) and cumulative distribution function (CDF), Monty Hall, game theory.

## II. Random Variables and Probability Distributions

1. Random variables, discrete random variables, continuous random variables, probability distribution, expectation, variance, standard deviation, infinite expectations, moments, functions of random variables.

2. Discrete random variables and several well-known discrete probability distributions: binomial, hypergeometric, geometric, negative binomial, poisson, and uniform, multinomial distributions.

3. Continuous random variables and several well-known continuous probability distributions: uniform, exponential, and normal distributions.
4. Joint Probability Distributions, random pairs, marginal distributions, independence, covariance, correlation coefficient, conditional expectation.

5. Chernoff bound, Markov and Chebyshev's inequalities, singular value decomposition (SVD), principle component analysis (PCA).

## III. Fundamental Sampling Distributions and Data Descriptions

1. Sampling distributions, sampling distribution of means and the Central Limit Theorem, sampling distribution of $S^2$

2. $t$-Distribution and F-Distribution

## IV. One- and Two-Sample Estimation Problems

1. Introduction, statistical inference, classical methods of estimation
2. Single sample: estimating the mean
3. Standard error of a point estimate
4. Properties of point estimators
5. Interval estimation: methods and properties of confidence interval
6. Sample size determination
7. Two samples: estimating the difference between two means
8. Paired observations
9. Single sample: estimating a proportion
10. Two samples: estimating the difference between two proportions
11. Single sample: estimating the variance
12. Two samples: estimating the ratio of two variances
13. Maximum likelihood estimation

## V. One- and Two-Sample Tests of Hypotheses

1. Statistical hypotheses: general concepts
2. Testing a statistical hypothesis
3. The use of p-values for decision making in testing hypotheses
4. Single sample: tests concerning a single mean
5. Two samples: tests on two means
6. Choice of sample size for testing means
7. One sample: test on a single proportion
8. Two samples: tests on two proportions
9. One- and two-sample tests concerning variances
10. Goodness-of-fit test
11. Test for independence (categorical data)

## VI. Simple Linear Regression and Correlation

1. Introduction to linear regression
2. The simple linear regression model
3. Least squares and the fitted model
4. Properties of the least squares estimators
5. Inferences concerning the regression coefficients
6. Prediction
7. Choice of a regression model
8. Analysis-of-variance approach
9. Test for linearity of regression: data with repeated observations

10. Simple linear regression case study

11. Correlation

## VII. Multiple Linear Regression and Certain Nonlinear Regression Models Estimating the Coefficients

1. Introduction
2. Linear regression model using matrices
3. Properties of the least squares estimators
4. Inferences in multiple linear regression
5. Choice of a fitted model through hypothesis testing
6. Categorical or indicator variables

## VIII. Time series

1. Secular trend (or general trend)
2. Seasonal movements
3. Cyclical movements
4. Irregular fluctuations

**Additional topics:**

## IX. Factor Experiments: General

1. Analysis-of-Variance Technique
2. The Strategy of Experimental Design
3. One-Way Analysis of Variance: Completely Randomized Design (One-Way ANOVA)

## X. **Nonparametric Statistics**

1. Nonparametric Tests
2. Signed-Rank Test
3. Wilcoxon Rank-Sum Test
4. Kruskal-Wallis Test

## XI. **Bayesian Statistics**

1. Bayesian Concepts
2. Bayesian Inferences

## Textbooks:

**Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

**Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber

**Elementary Statistics,** 13th Edition, Mario F. Triola

## Reference books:

**Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman

**Probability Demystified**, Allan G. Bluman

**Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson

**Python for Probability, Statistics, and Machine Learning,** José Unpingco

**Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce

**Think Stats: Probability and Statistics for Programmers,** Allen Downey

## Assessment types used:

A mix of weekly problem sets and programming assignments to be completed on Jupyter notebooks. Students are expected to have a computer on which they can install and run Jupyter.

**Distribution of marks:**

Midterm = 30 points

Final term = 40 points

Sessional points = 30 points

I. Assignments = 2 × 4 = 8 points

II. Hands-on Python in class = 0.5 × 6 =3 points

III. Quizzes = 2 × 6 = 12 points

IV. Journal/conference paper presentation = 5 points

V. Mini project (its report should be in an IEEE journal paper format) = 2 points

Or

The weightage of the project will be increased up to 10 points