This course covers both Descriptive statistics to understand the data and the inferential statistics which seeks to infer something about a population on the basis of a statistical sample and build a simple linear regression model. This course will also cover engineering optimization techniques like unconstrained, constrained, and discrete optimization techniques.

Course Objectives:
- Provide insights about the basic roles of a Data Scientist. Develop a greater understanding of the importance of Data Visualization techniques and Random Variables and their Distributions.
- Provide students with knowledge of Confidence Intervals and their importance. Make inferences about the population parameters using sample data and test it to draw meaningful conclusions.
- Provide an understanding on the importance and techniques of predicting a relationship between the two sets of data and determine the goodness of fit model.
- Provide an understanding on the importance and techniques of engineering optimization.

Course Outcomes:
At the end of this course, the student will be able to,
- Use Python and other tools to extract, clean and analyze data from several data sources (files, web) analyze an extremely large dataset and perform exploratory data analysis to extract meaningful insights.
- Analyze a real-world problem and solve the same with the knowledge gained from various distribution studies.
- Develop and test a hypothesis about the population parameters to draw meaningful conclusions and fit a regression model to data and use it for prediction.
- Analyze a real-world problem and solve the same with the knowledge gained from various engineering optimization techniques.

Course Content

Unit 1: Applications of Probability Distributions and Principles of Point Estimation
Introduction, Motivating Examples and Scope. Statistics: Introduction, Types of Statistics, Types of Data, Types of Experiments – Controlled and Observational study, Sampling: Sampling Methods, Sampling Errors, Case Study. Chebyshev's inequality, Normal Probability Plots, Introduction to Generation of Random Variates and mention the types, Acceptance-Rejection method, Sampling Distribution, The Central Limit Theorem and Applications, Principles of Point Estimation - Mean Squared Error for Bernoulli, Binomial, Poisson, Normal, Maximum Likelihood Estimate for Bernoulli, Binomial, Poisson, Normal and Case Study. Introduction to multivariate normal distribution, MAP distribution.
Self-Learning: Generation of Random Variates -Inverse Transform Method.

Applications:
1. Poisson distribution, calculation of number of calls received in a specified time duration in call centers.
2. Variance, standard deviation, identifying the customer satisfaction in online shopping
3. Central limit theorem, Load Balancing in distributed systems and internet traffic prediction
4. Sampling mean, Estimating database query response times

16 Hours

Unit 2: Confidence Intervals and Hypothesis Testing
Confidence Intervals: Interval Estimates for Mean of Large and Small Samples, Student's t Distribution, Interval Estimates for Proportion of Large and Small Samples, Confidence Intervals for the Difference between Two Means, Interval Estimates for Paired Data. Factors affecting Margin of Error, Hypothesis Testing for Population Mean and Population Proportion of Large

and Small Samples, Drawing conclusions from the results of Hypothesis tests, Case Study.
  Self-Learning: Confidence interval for difference between two proportions.
Applications:
1. t-distribution, confidence interval, students' performance analysis based on hours of study
2. z-test, application form processing in banking system.
3. Hypothesis testing, randomly trained students placement into tier-I and tier-II companies.

12 Hours

Unit 3: Distribution Free Tests and Multiple Linear Regression
Distribution Free Tests, Chi-squared Test, Fixed Level Testing, Type I and Type II Errors, Power of a Test, Factors Affecting Power of a Test. Simple Linear Regression: Introduction, Correlation, the Least Square Lines, Predictions using regression models - Uncertainties in Regression Coefficients, Checking Assumptions and transforming data, Introduction to the Multiple Regression Model, Case Study.
Self-Learning: F test for equality of Variance.

Applications:
1. Linear regression, stock market prediction
2. using Chi-Square Test, Analyzing the association between vaccination and recovery of the patients considering COVID data.
3. Chi-Square Test and Test of Independence, Analyzing the relationship between gender and preference for a product purchase.
4. Identifying Type 1 and Type 2 Errors in Spam mail classification.

14 Hours

Unit 4: Engineering optimization
Introduction to Optimization-Based Design, Modelling Concepts, Unconstrained Optimization, Discrete Variable Optimization, Genetic and Evolutionary Optimization, Constrained Optimization.
Self-Learning: Mathematical concepts of objective function, Constraints and Decision variables.

Applications:
1. Minimize a Loss functions in Neural Networks using Batch gradient descent (Unconstrained Optimization)
2. Lagrange Multipliers to find local maxima and minima of a function subject to equations constrains (Constrained Optimization)
3. Case study on Bayesian Optimization with Discrete Variables (Discrete Variable optimization)
4. Use Genetic Algorithms to optimize Production Scheduling in a manufacturing environment, focusing on minimizing total production costs while meeting job deadlines and machine constraints. Evaluate the GA's effectiveness against traditional scheduling methods.

14 Hours

Tools / Languages/Libraries: Jupyter Notebook**,** Python, Pandas, Matplotlib, Scipy, Seaborn, BeautifulSoup, Numpy, Scikit learn.

Text Book(s):
1. "Statistics for Engineers and Scientists", William Navidi, McGraw Hill Education, India, 4th Edition, 2015.
2. "Optimization Methods for Engineering Design, Parkinson, A.R., Balling, R., and J.D. Hedengren, Second Edition, Brigham Young University, 2018.

Reference Book(s):
1. "The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modelling", Raj Jain, Wiley, 2008.
2. "Sampling- Design and Analysis", Sharon L. Lohr, 2nd edition (stats), Cengage, 2010.
3. "Data Science from Scratch", Joel Grus, O"Reilly, 1st Edition, 2015.
4. "Statistics for Engineers and Scientists", William Navidi, 3rd Edition, McGraw Hill, 2010.