Course Code	Course Name	Credit
CSDLO5011	Statistics for Artificial IntelligenceData Science	03

Prerequisite: C Programming	
Course Objectives: The course aims:	
1	To Perform exploratory analysis on the datasets
2	To Understand the various distribution and sampling
3	To Perform Hypothesis Testing on datasets
4	To Explore different techniques for Summarizing Data
5	To Perform The Analysis of Variance
6	To Explore Linear Least Squares
Course Outcomes: Learner will be able to	
1	Illustrate Exploratory Data Analysis
2	Describe Data and Sampling Distributions
3	Solve Statistical Experiments and Significance Testing
4	Demonstrate Summarizing Data
5	Interpret the Analysis of Variance
6	Use Linear Least Squares

Prerequisite: Discrete Structures and Graph Theory

Module		Detailed Content	Hours
1		**Exploratory Data Analysis**	5
	1.1	Elements of Structured Data ,Further Reading ,Rectangular Data ,Data Frames and Indexes ,Nonrectangular Data Structures , Estimates of Location ,Mean ,Median and Robust Estimates , Estimates of Variability,Standard Deviation and Related Estimates ,Estimates Based on Percentiles , Exploring the Data Distribution ,Percentiles and Boxplots ,Frequency Tables and Histograms ,Density Plots and Estimates.	
	1.2	Exploring Binary and Categorical Data , Mode ,Expected Value, Probability ,Correlation ,Scatterplots ,Exploring Two or More Variables ,Hexagonal Binning and Contours (Plotting Numeric Versus Numerical Data) ,Two Categorical Variables ,Categorical and Numeric Data ,Visualizing Multiple Variables.	
2		**Data and Sampling Distributions**	6
	2.1	Random Sampling and Sample Bias ,Bias ,Random Selection ,Size Versus Quality,Sample Mean Versus Population Mean ,Selection Bias ,Regression to the Mean ,Sampling Distribution of a Statistic ,Central Limit Theorem ,Standard Error ,The Bootstrap ,Resampling Versus Bootstrapping .	
	2.2	Confidence Intervals ,Normal Distribution ,Standard Normal and QQ-Plots ,Long-Tailed Distributions ,Student's t-Distribution ,Binomial Distribution ,Chi-Square Distribution ,F-Distribution ,Poisson and Related Distributions ,Poisson Distributions ,Exponential Distribution ,Estimating the Failure Rate ,Weibull Distribution . **Self Study :** Problems in distributions.	
3		**Statistical Experiments and Significance Testing**	8
	3.1	A/B Testing ,Hypothesis Tests ,The Null Hypothesis ,Alternative Hypothesis ,One-Way Versus Two-Way Hypothesis Tests ,Resampling ,Permutation Test ,Example: Web Stickiness,Exhaustive and Bootstrap Permutation Tests ,Permutation Tests: The Bottom Line for Data Science ,Statistical Significance and p-Values ,p-Value ,Alpha ,Type 1 and	

		Type 2 Errors	
	3.2	Data Science and p-Values , t-Tests ,Multiple Testing ,Degrees of Freedom ,ANOVA ,F-Statistic ,Two-Way ANOVA , Chi-Square Test ,Chi-Square Test: A Resampling Approach ,Chi-Square Test: Statistical Theory ,Fisher's Exact Test ,Relevance for Data Science ,Multi-Arm Bandit Algorithm ,Power and Sample Size ,Sample Size . **Self Study :** Testing of Hypothesis using any statistical tool	
4		**Summarizing Data**	6
	4.1	Methods Based on the Cumulative Distribution Function , The Empirical Cumulative Distribution Function ,The Survival Function ,Quantile-Quantile Plots , Histograms, Density Curves, and Stem-and-Leaf Plots , Measures of Location.	
	4.2	The Arithmetic Mean ,The Median , The Trimmed Mean , M Estimates , Comparison of Location Estimates ,Estimating Variability of Location Estimates by the Bootstrap , Measures of Dispersion , Boxplots , Exploring Relationships with Scatterplots . **Self Study :** using any statistical tool perform data summarization	
5		**The Analysis of Variance**	6
	5.1	The One-Way Layout, Normal Theory; the F Test ,The Problem of Multiple Comparisons , A Nonparametric Method—The Kruskal-Wallis Test ,The Two-Way Layout , Additive Parametrization , Normal Theory for the Two-Way Layout ,Randomized Block Designs , A Nonparametric Method—Friedman's Test .	
6		**Linear Least Squares**	8
	6.1	Simple Linear Regression, Statistical Properties of the Estimated Slope and Intercept , Assessing the Fit , Correlation and Regression , The Matrix Approach to Linear Least Squares , Statistical Properties of Least Squares Estimates , Vector-Valued Random Variables , Mean and Covariance of Least Squares Estimates , Estimation of $\sigma 2$, Residuals and Standardized Residuals , Inference about β , Multiple Linear Regression—An Example , Conditional Inference, Unconditional Inference, and the Bootstrap , Local Linear Smoothing . **Self Study :**Create a Linear Regression model for a dataset and display the error measures, Chose a dataset with categorical data and apply linear regression model	

Textbooks:	
1	Bruce, Peter, and Andrew Bruce. Practical statistics for data scientists: 50 essential concepts. Reilly Media, 2017.
2	Mathematical Statistics and Data Analysis John A. Rice University of California, Berkeley,Thomson Higher Education
References:	
1	Dodge, Yadolah, ed. Statistical data analysis and inference. Elsevier, 2014.
2	Ismay, Chester, and Albert Y. Kim. Statistical Inference via Data Science: A Modern Dive into R and the Tidyverse. CRC Press, 2019.
3	Milton. J. S. and Arnold. J.C., "Introduction to Probability and Statistics", Tata McGraw Hill, 4th Edition, 2007.
4	Johnson. R.A. and Gupta. C.B., "Miller and Freund's Probability and Statistics for Engineers", Pearson Education, Asia, 7th Edition, 2007.
5	A. Chandrasekaran, G. Kavitha, "Probability, Statistics, Random Processes and Queuing Theory", Dhanam Publications, 2014.

Assessment:	
Internal Assessment:	
Assessment consists of two class tests of 20 marks each. The first-class test is to be conducted when approx. 40% syllabus is completed and second class test when additional40% syllabus is completed. Duration of each test shall be one hour.	
End Semester Theory Examination:	
1	Question paper will consist of 6 questions, each carrying 20 marks.
2	The students need to solve a total of 4 questions.
3	Question No.1 will be compulsory and based on the entire syllabus.
4	Remaining question (Q.2 to Q.6) will be selected from all the modules.

Useful Links	
1	https://www.edx.org/course/introduction-probability-science-mitx-6-041x-2
2	https://www.coursera.org/learn/statistical-inference
3	https://www.datacamp.com/community/open-courses/statistical-inference-and-data-analysis

*** Suggestion: Laboratory work based on the above syllabus can be incorporated as a mini project in CSM501: Mini-Project.**