# INT351:STATISTICS AND EDA

L:2 T:0 P:3 Credits:4

**Course Outcomes:** Through this course students should be able to

CO1 :: understand the basic concepts of exploratory data analysis.

CO2 :: apply the univariate techniques on the data to analyze it.

CO3 :: apply the concepts of probability and distribution to depicts the expected outcomes of possible values for a given data.

CO4 :: study the different methods to explore the data for analysis

CO5 :: understand the different types of inferential statistics

CO6 :: study the different types of testing on data to perform the inferential statistics

**Unit I**

**EDA** : introduction to EDA, public and private data, web scraping, data types, fixing the rows and columns, impute/remove missing values, handling outliers, standardizing values, fixing invalid values and filter data

**Unit II**

**Univariate analysis** : introduction to univariate analysis, categorical unordered univariate analysis, categorical ordered univariate analysis, statistics on numerical features, numeric - numeric analysis, correlation vs causation, numerical - categorical analysis, categorical - categorical analysis, multivariate analysis, missing values, correlation and causation, data visualization

**Unit III**

**Inferential statistics-I** : permutations, combinations, probability: definition and properties, types of events, rules of probability - addition, rules of probability - multiplication, introduction: basics of probability, random variables, probability distributions, expected value, summary - basics of probability, introduction: discrete probability distributions, binomial distribution, binomial distribution (examples), cumulative probability, comprehension: expected value, introduction: continuous probability distributions, probability density functions, normal distribution, standard normal distribution

**Unit IV**

**Inferential statistics-II** : introduction, central limit theorem, samples, sampling distributions, properties of sampling distributions, clt - demonstration, estimating mean using clt, confidence interval - example, types of sampling methods, uses of sampling in market research, uses of sampling in market campaigns, uses of sampling in pilot testing, uses of sampling in quality control, basics of probability, joint probability and conditional probability, bayes' theorem, standardized normal distribution and z - score, sampling methods, sampling and estimation

**Unit V**

**Hypothesis testing** : understanding hypothesis testing, null and alternative hypotheses, making a decision, critical value method, critical value method examples, the p-value method, the p-value method: examples, types of errors

**Unit VI**

**Hypothesis formulation** : introduction, choosing the representative sample, computing the test-statistic, finding the critical region, making the decision, using p-value approach, changing the hypothesis, t distribution, two-sample mean test, two-sample proportion test, a/b testing demonstration, industry relevance, hypothesis testing in python, distributions and sampling methods, inferential statistics, hypothesis testing, a/b testing, chi-squared test, ANOVA

**List of Practicals / Experiments:**

**List of practical**

- demonstration of importing dataset from various sources to the jupyter notebook
- analysis of quantity and distribution of data and understanding it by interpreting the columns
- demonstrating the use descriptions or comments to describe the finding and storytelling
- analysis of data quality issues and perform various techniques for missing value imputation
- understanding the data outliers and remove duplicates and irrelevant data from the dataset

- transforming and preparing the data with the right data types using conversion function
- performing univariate analysis
- performing categorical unordered univariate and ordered univariate analysis
- performing bivariate analysis and multivariate analysis
- constructing hypothesis around the business problem
- understanding discrete probability distributions and demonstrating its applications
- understanding continuous probability distributions and demonstrating its applications
- applying the concepts of central limit theorem on the dataset
- performing hypothesis testing using critical value and p-value methods

**References:**

1. STATISTICAL LEARNING AND DATA SCIENCE by MIREILLE GETTLER SUMMA, LEON BOTTOU, BERNARD GOLDFARB, FIONN MURTAGH, CATHERINE PARDOUX, MYRIAM TOUATI, CHAPMAN & HALL (CRC PRESS)

2. STATISTICAL ANALYSIS by DR. B. N. GUPTA, SBPD Publications

3. HANDS-ON EXPLORATORY DATA ANALYSIS WITH PYTHON: PERFORM EDA TECHNIQUES TO UNDERSTAND, SUMMARIZE, AND INVESTIGATE YOUR DATA by SURESH KUMAR MUKHIYA , USMAN AHMED, PACKT PUBLISHING