Statistics For Data Science

Statistics is a broad field and is used in many industries. Most Corporate Management use statistics in what is called as “Decision science”. Multiple Government agencies also use statistics while planning to spend resources for common good.

Wikipedia defines Statistics as: Statistics is the collection, analysis, interpretation, presentation and organisation of data. So, it’s not surprising that individuals aspiring to develop career in the field of Data, need to understand Statistics.

This course aims at bring students up to speed with advanced statistics required for building a career in the field of Data science.

Need to learn Statistics

- The psychology of Statistics.
- The cautionary tale of Simpson’s Paradox.
- Statistics in everyday life.
- Scales of Measurement
- Accessing the reliability of a Measure.
- The Role of variables: Predictors & Outcomes.

Descriptive Statistics

- Measures of Central Tendencies.
- Measures of Variation.
- Skew & Kurtosis.
- Getting the overall summary of a variable.
- Descriptive Statistics separately for each group.
- Standard Scores
- Correlations
- Handling Missing values.

Introduction to Probability

- How is probability and statistics different?
- The law of large numbers.
- Sampling distribution and the law of large numbers.
- Estimating population parameters
- Estimating a confidence interval.

Hypothesis Testing

- The world of Hypothesis.
- The two types of Error.
- Test Statistics and Sampling distribution.
- Making decisions
- The p-value of a test.
- Reporting the results of a Hypothesis test.

- Running the Hypothesis test in practice.
- Effect size, sample size and power.

Categorical Data Analysis

- The x^2 Goodness of fit test.
- The x^2 test of independence (or association).
- The Continuity correction.
- Effect size
- Assumption of the tests.
- Chi-Square test in R
- The Fishers exact test.
- The McNemar Test.
- What is the difference between McNemar & Independence?

Comparing Two Means

- The one sample Z test.
- The one sample t test.
- The independent samples t-test (student test).
- The independent samples t-test (welch test)
- The paired samples t-test.
- One sided test.
- Using the t.test() function.
- Effect size
- Checking the normality of sample.
- Testing the non normal data Wilcoxon test.

Comparing Several Means (One-Way ANOVA)

- How ANOVA Works.
- Running ANOVA in R.
- Effect size.
- Multiple comparisons and post hoc tests.
- Assumptions of One Way ANOVA.
- Checking Homogeneity of Variance assumption.
- Removing Homogeneity of variance assumption.
- Checking Normality assumption.
- Removing Normality assumption.
- Relationship between ANOVA and the student t test.

Linear Regression

- What is a linear Regression Model?
- Estimating the linear Regression Model.
- Multiple Linear Regression.
- Quantifying the fit of the regression model.
- Hypothesis test for the Regression Model.
- Testing the significance of a correlation.
- Regarding Regression coefficients.

- Model checking.
- Assumptions of Regression.
- Model checking.
- Model selection.

Factorial ANOVA

- Factorial ANOVA 1: Balanced design, no interaction.
- Factorial ANOVA 2: Balanced design, interaction allowed.
- Effect size, estimated means, and confidence intervals.
- Assumptions checking.
- The F-test as a model comparison.
- ANOVA as a linear model.
- Different ways to specify contrasts.
- Post hoc tests.
- The method of planned comparisons.
- Factorial ANOVA 3: Unbalanced designs.

Gateway to Machine learning

- How Statistics and Machine learning are different yet complimentary.