IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Basic Statistics Using R

LECTURE 04

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

Basic Statistics

- Descriptive Statistics
  - Open RStudio
- Hypothesis Testing
  - Formulate an assertion and test it using data
    - Comparing populations, e.g., comparing performance of students in exams for two different class sections
    - Testing the difference of the means from two data samples
  - A common technique to assess the difference or significance of the same

Basic Statistics

- Common assumption in Hypothesis testing
  - No difference between two samples
  - Referred as NULL Hypothesis H_0
  - Alternative Hypothesis (H_A): There is difference between two samples
- Example:
  - H_0: Students from class A and B had same performance in the examinations
  - H_A: Students from class A performed better than students from class B

Basic Statistics

- Hypothesis test leads to:
  - Either rejection of the null hypothesis in favor of the alternative
  - Or acceptance of the null hypothesis
- Examples:
  - H_0: New data mining model does not predict better than existing model
  - H_A: New data mining model predicts better than existing model

Basic Statistics

- Examples:
  - H_0: Regression coefficient is zero, i.e., variable has no impact on outcome
  - H_A: Regression coefficient is nonzero, i.e., variable has an impact on outcome
- A typical hypothesis test is comparing the means of two populations
- Normal Distribution
  - A common continuous probability distribution and useful due to Central limit theorem

Basic Statistics

- Difference of Means
  - Drawing inferences on two populations: P1 and P2
  - Compare means: μ_1 and μ_2
  - H_0: $\mu_1 = \mu_2$
  - H_A: $\mu_1 \neq \mu_2$
  - Basic approach: compare observed sample means: $\bar{x}_1$ and $\bar{x}_2$
- Student's t-test
  - Assumptions: Two population distributions (P1 and P2) have equal but unknown variances
  - Two samples of n_1 and n_2 observations drawn randomly and independently from P1 and P2, respectively

Basic Statistics

- Student's t-test
  - If P1 and P2 are normally distributed with same mean and variance
  - Then t-statistic follows a t-distribution with n_1+n_2-2 degrees of freedom

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{Where } {S_p}^2 = \frac{(n_1-1){S_1}^2+(n_2-1){S_2}^2}{n_1+n_2-2}$$

Basic Statistics

- Student's t-test
  - S_p is pooled standard deviation, S_1 and S_2 are sample standard deviation
  - Shape of t-distribution is similar to normal distribution and becomes identical to normal distribution as degrees of freedom reach 30 or more
  - Numerator of t is the difference of the sample means
    - Observed t value of 0 indicates the sample results are exactly equal to H_0
    - Observed t value being far enough from 0 and t-distribution indicating a low enough probability (<0.05) will lead to rejection of H_0
    - t-value falling in corresponding areas in the curve less than 5% of the time

Thanks…