IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

NPTEL

Basic Statistics Using R Part-2

LECTURE 05

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

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

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

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

Basic Statistics

- Student's t-test
  - For a low probability, $\alpha = 0.05$, known as significance level of the test
  - t^* is determined such that $p(|t| \geq t^*) = \alpha$
  - H_0 is rejected if observed value of t is such that $|t| \geq t^*$
- Significance level of a statistical test is the probability of rejecting the null hypothesis
  - If null hypothesis is true and $\alpha = 0.05$, the observed magnitude of t would exceed t^* 5% of the time

Basic Statistics

- p-value is sum of p(t≤-|observed t-value|) and p(t≥|observed t-value|)
- Open Rstudio
- Welch's t-test
  - Used when assumption of equal population variance is not reasonable

$$t_w = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{S_1}^2}{n_1} + \frac{{S_2}^2}{n_2}}}$$

Basic Statistics

- Welch's t-test
  - Assumption of random samples drawn from two normal populations with the same mean is still applicable
  - t-distribution
- Open RStudio
- Confidence Interval
  - Provide interval estimate of a population parameter using sample data
  - Indicates uncertainty associated with a point estimate
  - How close $\bar{x}$ is to μ

Basic Statistics

- Confidence Interval
  - A 95% confidence interval estimate for a population mean straddles the true unknown mean 95% of the time

$$\mu \in \bar{x} \pm \frac{2\sigma}{\sqrt{n}}$$

- Type I and Type II Errors

	H_0 is true	**H_0 is false**
H_0 accepted		Type II error
H_0 rejected	Type I error	

Basic Statistics

- Type I and Type II Errors
  - Significance level = type I error (Denoted by α)
    - Can be managed using appropriate significance level
  - Type II error (Denoted by β)
    - Can be managed using appropriate sample size
- Power of a test
  - Correctly rejecting H_0
  - $1- \beta$
  - Used to determine the sample size

Basic Statistics

- ANOVA
  - Used for more than two populations or groups instead of performing multiple t-tests
  - Generalization of hypothesis testing that is used for the difference of two group means
  - For n groups, n(n-1)/2 t-tests would be required
  - Multiple t-tests
    - Cognitively difficult
    - Increased probability of type I error

Basic Statistics

- ANOVA
  - H_0: All the population means are equal
  - H_A: At least one pair of the population means is not equal
  - Assumption: Each population is normally distributed with same variance
  - Test whether different population clusters are more tightly grouped or spread across all the populations

Basic Statistics

- ANOVA
  - Between-groups mean sum of squares ($S_B{}^2$)
    - An estimate of between-groups variance

$$S_B{}^2 = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_0)^2$$

Where k=no. of groups, n_i is no. of observations in ith group, $\bar{x}_0$ is mean of all the groups, $\bar{x}_i$ is mean of ith group

  - Within-group mean sum of squares ($S_W{}^2$)
    - An estimate of within-group variance

Basic Statistics

- ANOVA
  - Within-group mean sum of squares (S_W^2)

$$S_W{}^2 = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} n_i(x_{ij} - \bar{x}_i)^2$$

  - If $S_B^2 > S_W^2$, some of the population means are different
  - F-test statistic

$$F = \frac{S_B{}^2}{S_W{}^2}$$

- Open RStudio

Thanks…