Lecture 25

Confidence Intervals (for difference between 2 groups)

Manju M. Johny

STAT 330 - Iowa State University

Confidence Interval for Difference Between Groups

Confidence Intervals

- In the previous lecture, we learned how to build a confidence interval to estimate an unknown population parameter θ

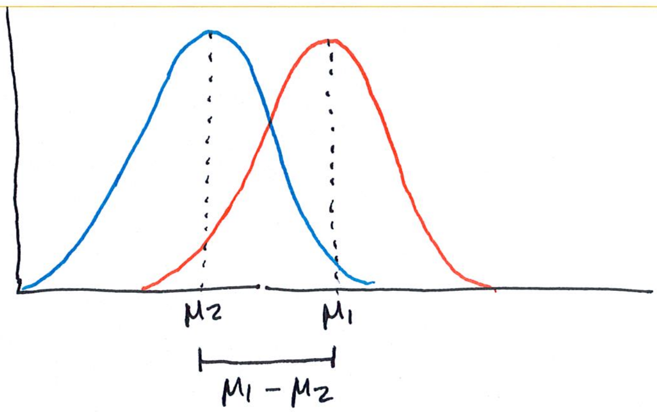
$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

- Now, we learn how to build a confidence interval to estimate the *difference* between 2 population parameters
  - → Compare group 1 and group 2 with parameters θ_1 and θ_2 respectively
  - → Build a confidence interval for unknown $\theta_1 - \theta_2$

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2} SE(\hat{\theta}_1 - \hat{\theta}_2)$$

CI for Difference in Means

CI for Difference Between Means ($\mu_1 - \mu_2$)



- Group 1 has unknown population mean μ_1
- Group 2 has unknown population mean μ_2
- Build a confidence interval to estimate $\mu_1 - \mu_2$

CI for $\mu_1 - \mu_2$ Cont.

Estimate $\mu_1 - \mu_2$ with $(\bar{X}_1 - \bar{X}_2)$

- $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$
- $Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

  $\rightarrow SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Since we typically don't know the population variance σ^2, replace it with the sample variance s^2.

$\rightarrow SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

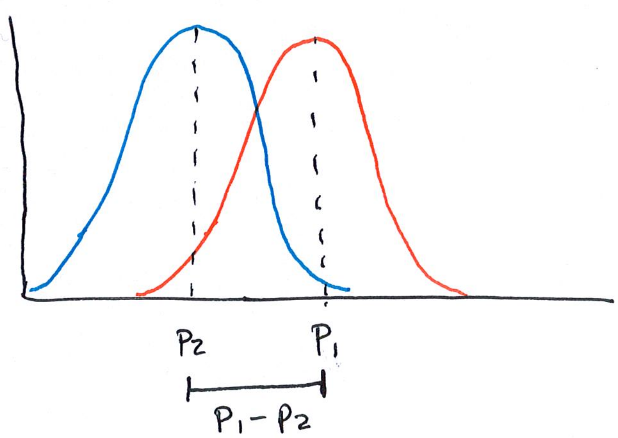
Then, the confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Formula sheet

CI for Difference in Proportions

CI for Difference Between Proportions $(p_1 - p_2)$



- Group 1 has unknown population proportion p_1
- Group 2 has unknown population proportion p_2
- Build a confidence interval to estimate $p_1 - p_2$

CI for $p_1 - p_2$ Cont.

Estimate $(p_1 - p_2)$ with $\hat{p}_1 - \hat{p}_2$

- $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
- $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$

  $\rightarrow SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Since we don't know the population proportion p, replace it with sample proportion $\hat{p}$.

$\rightarrow SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

Then, the confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Formula Sheet

Examples

Example: Difference in Means

Example 1: Taxable Income

Group 1 (East)
$n_1 = 1000$
$\bar{x}_1 = 37200$
$s_1 = 10100$

Group 2 (West)
$n_2 = 2000$
$\bar{x}_2 = 42000$
$s_2 = 15,600$

95% confidence
$\rightarrow z_{\alpha/2} = 1.96$

We obtain IRS records from the east coast and the west coast for the year 2000. For 1000 records obtained from the east coast, the mean taxable income is \$37, 200 and standard deviation is \$10, 100. For 2000 records obtained from the west coast, the mean taxable income is \$42, 000 and standard deviation is \$15, 600. Construct a 95% confidence interval to compared the mean taxable income between the 2 regions.

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (37200 - 42000) \pm 1.96\sqrt{\frac{10100^2}{1000} + \frac{15600^2}{2000}}$$

$$= -4800 \pm 927$$

$$= [-5726, -3873]$$

$$\underbrace{[-5726 \quad\quad\quad -3873]}_{\text{reasonable values for } \mu_1 - \mu_2}$$

Since both ends of my CI are negative, we think group 1 − group 2 < 0
$\Rightarrow$ East − West < 0
$\Rightarrow$ East < West

Interpretation

- We are 95% confident that the difference in population mean taxable incomes between east coast and west coast (east - west) is between -\$5, 726 and -\$3, 873

  OR

- We are 95% confident that the true mean taxable income in the east coast is *less* than that of the west cost by between \$3, 873 and \$5, 726.

Example: Difference in Proportions

Example 2: Digital Communications

Group 1 (A)

$n_1 = 100$

$\hat{p}_1 = \frac{5}{100} = 0.05$

Group 2 (B)

$n_2 = 100$

$\hat{p}_2 = \frac{10}{100} = 0.10$

98% confidence

$\rightarrow z_{\alpha/2} = 2.326$

Suppose we are interested in comparing the corruption rates of messages sent using 2 different digital communication systems. Out of a 100 messages sent by system A, 5 are corrupted in transmission. Out of a 100 messages sent by system B, 10 are corrupted in transmission. What's the difference in the corruption rates? Calculate a 98% confidence interval to estimate the difference in the corruption rates.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\Rightarrow (0.05 - 0.10) \pm 2.326\sqrt{\frac{0.05(0.95)}{100} + \frac{0.10(0.90)}{100}}$$

$$\Rightarrow -0.05 \pm 0.086$$

$$\Rightarrow [-0.136, 0.036]$$

lower: -0.136

group1 − group2 < 0

$\Rightarrow A < B$

upper: 0.036

group1 − group > 0

$\Rightarrow A > B$

$[$ −0.136 ———— 0.036 $]$

$\underbrace{-0.136 \quad \quad 0.036}_{\text{reasonable values for } p_1 - p_2}$

Interpretation

- We are 98% confident that the difference in true corruption rates between system A and B (A - B) is between -0.136 and 0.036.

  OR

- We are 98% confident that the population corruption rate of system A is between 0.136 *less* than and 0.036 *greater* than the population corruption rate of system B.

Note: Since 0 is contained in the confidence interval, there is no significant evidence of difference between system A and B.