ISyE 6739 – Statistical Methods

Confidence Intervals – Two Populations

Instructor: Kamran Paynabar
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

kamran.paynabar@isye.gatech.edu
Office: Groseclose 436

Outline

CI for Difference in Means

- Normally distributed data
  - with known σ
  - with unknown σ
- Non-normal data with large sample size (n>30)
  - with known/unknown σ

CI for Ratio of Variances

- Normally distributed data

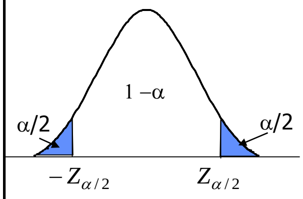
CI for Difference of Proportions

- Large sample size

100(1-α)% CI for Differences in Means of Two Normal Distributions (two-sided, known variances)

Two independent random samples from two Normal distributions with the known variances

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim NID(\mu_1, \sigma_1^2) \\ \\ Y_1, Y_2, \cdots, Y_{n_2} \sim NID(\mu_2, \sigma_2^2) \end{array}\right\} \rightarrow \bar{X} - \bar{Y} \sim NID(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim NID(0,1)$$



$$\Pr\left\{ -Z_{\alpha/2} \le \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z_{\alpha/2} \right\} = 1 - \alpha \longleftarrow$$ **Confidence level**

CI Lower **CI Upper**

$$\boxed{(\bar{X} - \bar{Y}) - Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le \mu_1 - \mu_2 \le \boxed{(\bar{X} - \bar{Y}) + Z_{\alpha/2} \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

This CI can be used for mean of non-normal distributions when n>30

Example

Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10$, $\bar{x}_1 = 87.6$, $\sigma_1 = 1$, $n_2 = 12$, $\bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

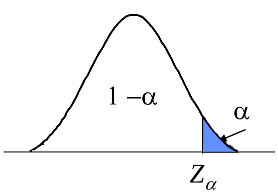
Choice of Sample Size

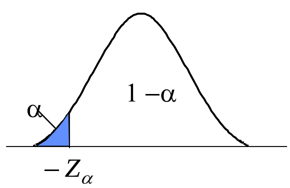
$$Error = \left|(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)\right|$$

Find sample size (n) such that, Error < E with 1-α confidence.

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 (\sigma_1^2 + \sigma_2^2) \qquad (10\text{-}8)$$

100(1-α)% CI for Differences in Means of Two Normal Distributions (one-sided, known variances)





$$\Pr\left\{\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le Z_\alpha\right\} = 1 - \alpha$$

$$\Downarrow$$

$$\boxed{(\mu_1 - \mu_2) \ge (\bar{X} - \bar{Y}) - Z_\alpha \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Lower bound

$$\Pr\left\{-Z_\alpha \le \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right\} = 1 - \alpha$$

$$\Downarrow$$

$$\boxed{(\mu_1 - \mu_2) \le (\bar{X} - \bar{Y}) + Z_\alpha \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Upper bound

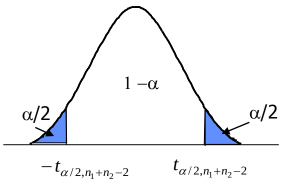
This CI can be used for mean of non-normal distributions when n>30

100(1-α)% CI for Difference in Means of Two Normal Distributions (two-sided, unknown variances)

Case I: Two independent random samples from two Normal distributions with unknown (but equal) variances $\sigma_1^2 = \sigma_2^2 = \sigma = ?$

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim NID(\mu_1, \sigma_1^2) \\ \\ Y_1, Y_2, \cdots, Y_{n_2} \sim NID(\mu_2, \sigma_2^2) \end{array}\right\} \rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

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



$$P\left\{ -t_{\alpha/2, n_1+n_2-2} \le \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \le t_{\alpha/2, n_1+n_2-2} \right\} = 1 - \alpha$$

$$\boxed{(\bar{X} - \bar{Y}) - t_{\alpha/2, n_1+n_2-2} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X} - \bar{Y}) + t_{\alpha/2, n_1+n_2-2} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Note that $t_{\alpha/2, n_1+n_2-2} \approx Z_{\alpha/2}; \text{for } n > 30$

Example

An article in the journal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$, with a sample standard deviation of $s_1 = 5.0$, while 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$, with a sample standard deviation of $s_2 = 4.0$.

$\sigma_1^2 = \sigma_2^2 = \sigma = ?$

100(1-α)% CI for Difference in Means of Two Normal Distributions (one-sided, unknown variances)

Case I: Two independent random samples from two Normal distributions with unknown (but equal) variances $\sigma_1^2 = \sigma_2^2 = \sigma = ?$

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim NID(\mu_1, \sigma_1^2) \\ Y_1, Y_2, \cdots, Y_{n_2} \sim NID(\mu_2, \sigma_2^2) \end{array}\right\} \rightarrow \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

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

$$\boxed{(\bar{X} - \bar{Y}) - t_{\alpha, n_1+n_2-2} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2}$$

Lower bound

$$\boxed{\mu_1 - \mu_2 \le (\bar{X} - \bar{Y}) + t_{\alpha, n_1+n_2-2} \times S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Upper bound

Note that $t_{\alpha/2, n_1+n_2-2} \approx Z_{\alpha/2}; \text{for } n > 30$

100(1-α)% CI for Difference in Means of Two Normal Distributions (unknown variances)

Case II: Two independent random samples from two Normal distributions with unknown and not equal variances $\sigma_1^2 \ne \sigma_2^2$

Two-sided

$$\boxed{(\bar{X} - \bar{Y}) - t_{\alpha/2, v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{X} - \bar{Y}) + t_{\alpha/2, v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

One-sided

$$\boxed{(\bar{X} - \bar{Y}) - t_{\alpha, v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2}$$

Lower bound

$$\boxed{\mu_1 - \mu_2 \le (\bar{X} - \bar{Y}) + t_{\alpha, v} \times \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Upper bound

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Note that $t_{\alpha/2, n_1+n_2-2} \approx Z_{\alpha/2}; \text{for } n > 30$

Example

An article in the journal *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) reported the results of an analysis of the weight of calcium in standard cement and cement doped with lead. Reduced levels of calcium would indicate that the hydration mechanism in the cement is blocked and would allow water to attack various locations in the cement structure. Ten samples of standard cement had an average weight percent calcium of $\bar{x}_1 = 90.0$, with a sample standard deviation of $s_1 = 5.0$, while 15 samples of the lead-doped cement had an average weight percent calcium of $\bar{x}_2 = 87.0$, with a sample standard deviation of $s_2 = 4.0$.

$\sigma_1^2 \neq \sigma_2^2$

100(1-α)% CI for Variances Ratio of Two Normal Distributions

Two independent random samples from two Normal distributions

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim NID(\mu_1, \sigma_1^2) \\ \\ Y_1, Y_2, \cdots, Y_{n_2} \sim NID(\mu_2, \sigma_2^2) \end{array}\right\} \rightarrow \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

$$\boxed{F_{1-\alpha, v_1, v_2} = \frac{1}{F_{\alpha, v_2, v_1}}}$$

$$\boxed{\frac{S_1^2}{S_2^2} F_{1-\alpha/2, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} F_{\alpha/2, n_2-1, n_1-1}}$$

Two-sided

$$\boxed{\frac{S_1^2}{S_2^2} F_{1-\alpha, n_2-1, n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2}}$$

Lower bound

$$\boxed{\frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2} F_{\alpha, n_2-1, n_1-1}}$$

Upper bound

One-sided

Example

A company manufactures impellers for use in jet-turbine engines. One of the operations involves grinding a particular surface finish on a titanium alloy component. Two different grinding processes can be used, and both processes can produce parts at identical mean surface roughness. The manufacturing engineer would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $s_1 = 5.1$ microinches, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $s_2 = 4.7$ microinches. We will find a 90% confidence interval on the ratio of the two standard deviations, σ_1/σ_2.

100(1-α)% CI for Difference in Population Proportions with large sample size

$$\left.\begin{array}{l} X_1, X_2, \cdots, X_{n_1} \sim Bernoulli(p_1) \\ Y_1, Y_2, \cdots, Y_{n_2} \sim Bernoulli(p_2) \end{array}\right\} \rightarrow \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

For large sample sizes

Two-sided CI

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le (p_1 - p_2) \le (\hat{p}_1 - \hat{p}_2) + Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) - Z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le (p_1 - p_2)$$

$$(p_1 - p_2) \le (\hat{p}_1 - \hat{p}_2) + Z_{\alpha}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Lower bound **Upper bound**

One-sided

Example

Consider the process manufacturing crankshaft bearings described in Example 8-6. Suppose that a modification is made in the surface finishing process and that, subsequently, a second random sample of 85 axle shafts is obtained. The number of defective shafts in this second sample is 8. Therefore, since $n_1 = 85$, $\hat{p}_1 = 0.12$, $n_2 = 85$, and $\hat{p}_2 = 8/85 = 0.09$, we can obtain an approximate 95% confidence interval on the difference in the proportion of defective bearings produced under the two processes