Stat 201: Introduction to Statistics

Standard 34: Confidence Intervals – for Mean Differences

Telling Which Parameter We're After

- As statisticians, or data scientists, it's our job to hear a problem and decide what we're after
  - We call the parameter of interest the **target parameter**

Parameter	Point Estimate	Key Phrase	Type of Data
$\mu_1 - \mu_2$	$\overline{x_1} - \overline{x_2}$	Mean Difference	Quantitative
$\rho_1 - \rho_2$	$\widehat{p_1} - \widehat{p_2}$	Difference of Proportion, percentage, fraction, rate	Qualitative (Categorical)

Note!

- We use the usual approach of confidence intervals and hypothesis testing on the mean difference $\mu_d = \mu_1 - \mu_2$ as we did in chapters 7-9
  - Our data becomes $x_{d_i} = x_{1_i} - x_{2_i}$ and we are interested in making inference on μ_d.

Now means…

- Just like before, we will transition from proportions to means.
- We will look at the difference of quantitative variables now – the difference of means.

Difference of Means

- We’re often interested in comparing groups of data.
- We follow similar steps from our Means confidence intervals
- We follow similar steps from our Hypothesis Testing for Means, but a couple of the formulas change

Difference of Means

- In the frame of Chapter 7, **sampling distributions**, we need to find the mean and standard deviation of **repeated samples** of mean differences.

Sampling Distribution for Mean Difference

- **The sample mean difference** is the sample mean of group 1 minus the sample mean of group 2
  - $\overline{x_d}$ = $\bar{x}_1$ - $\bar{x}_2$

Sampling Distribution for Mean Difference

- **The population mean of the sample mean differences** is the population mean of group 1 minus the population mean of group 2
  - $\mu_d = \mu_1 - \mu_2$

Sampling Distribution for Mean Difference

- **The standard error, or the standard deviation of all possible the sample mean differences,** is seen below:

- $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
  - where n_1 & n_2 are the number of people in each group and s_1^2 & s_2^2 are the sample variance for each group

Difference of Means

- In the frame of Chapter 8, **confidence intervals**, we need to find the **point estimate** and **margin of error** so that we can come up with an interval estimate of the **population mean difference**.

Confidence Intervals
Case One With known σ_1 & σ_2

- Check the assumptions
  1. Each sample must be obtained through randomization
  2. Samples are **independent**
  3. The differences are from the normal distribution
     - If $n_1 > 30$ & $n_2 > 30$

       <u>**OR**</u>
     - If both populations follow the normal distribution

Confidence Intervals
Case One With known σ_1 & σ_2

- We use our sample means to make inference on the population mean

$$(\overline{x_1} - \overline{x_2}) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- $\overline{x_1} - \overline{x_2}$ is our **point-estimate** for the population mean

- $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ is our **margin of error**

Confidence Intervals
Case One With known σ_1 & σ_2

- $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ is our **margin of error**
  - **As either n increases**, $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ decreases, causing the margin of error to decrease causing the width of the confidence interval to narrow
  - **As either n decreases**, $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ increases, causing the margin of error to increase causing the width of the confidence interval to widen

Confidence Intervals
Case One With known σ_1 & σ_2

- $z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ is our **margin of error**
  - **As the confidence level decreases**, z decreases causing the margin of error to decrease, causing the width of the confidence interval to narrow
  - **As the confidence level increases**, z increases causing the margin of error to increase, causing the width of the confidence interval to grow wider

Confidence Intervals
Case One With known σ_1 & σ_2

Lower Bound = $(\overline{x_1} - \overline{x_2}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Upper Bound =$(\overline{x_1} - \overline{x_2}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

We are --% confident that the true population difference of means is between the **lower** and **upper** bound.

Confidence Intervals
Case One With known σ_1 & σ_2

- $\bar{x}_{\mathrm{d}} = \bar{x}_1 - \bar{x}_2$
- Confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the resulting interval, (L,U), has both L and U less than 0 this suggests that the true mean difference, $\mu_d = \mu_1 - \mu_2$, is negative.
- $\mu_d = \mu_1 - \mu_2 < 0$ indicates that $\mu_1 < \mu_2$, that group 2 has the greater mean.

Confidence Intervals
Case One With known σ_1 & σ_2

- $\bar{x}_{\mathrm{d}} = \bar{x}_1 - \bar{x}_2$
- Confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the resulting interval, (L,U), has both L and U greater than 0 this suggests that the true mean difference, $\mu_d = \mu_1 - \mu_2$, is positive.
- $\mu_d = \mu_1 - \mu_2 > 0$ indicates that $\mu_1 > \mu_2$, that group 1 has the greater mean.

Confidence Intervals
Case One With known σ_1 & σ_2

- $\bar{x}_{\mathrm{d}} = \bar{x}_1 - \bar{x}_2$
- Confidence interval is given by:

$$\bar{x}_{\mathrm{d}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the resulting interval, (L,U), contains 0 this suggests that the true mean difference can be 0, $\mu_d = \mu_1 - \mu_2 = 0$.
- $\mu_d = \mu_1 - \mu_2 = 0$ indicates that $\mu_1 = \mu_2$, that the two groups can have the same mean

Confidence Intervals
Case One With known σ_1 & σ_2

- If all the values on the interval are negative then $\mu_1 < \mu_2$

- If all the values on the interval are positive then $\mu_1 > \mu_2$

- If 0 is on the interval then it's possible that $\mu_1 = \mu_2$

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

- Check the assumptions
  1. Each sample must be obtained through randomization
  2. Samples are **independent**
  3. The differences are from the normal distribution
     - If $n_1 > 30$ & $n_2 > 30$

       OR

     - If both populations follow the normal distribution

- **Note:** We will use StatCrunch to do all of these calculations.

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

- We use our sample means to make inference on the population mean

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\frac{\alpha}{2}, n_1+n_2-1} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where: $s_p^2 = \frac{(\mathrm{n}_1 - 1)\mathrm{s}_1^2 + (\mathrm{n}_2 - 1)\mathrm{s}_2^2}{n_1 + n_2 - 2}$

- $\overline{x_1} - \overline{x_2}$ is our **point-estimate** for the population mean
- $t_{1-\frac{\alpha}{2}, n_1+n_2-1} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ is our **margin of error**

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

- $t_{1-\frac{\alpha}{2},n_1+n_2-1}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ is our **margin of error**
  - **As either n increases**, $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ decreases, causing the margin of error to decrease causing the width of the confidence interval to narrow
  - **As either n decreases**, $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ increases, causing the margin of error to increase causing the width of the confidence interval to widen
  - **If one decreases and the other increases** – we have to plug in the values to see what the overall effect is

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

- $t_{1-\frac{\alpha}{2},n_1+n_2-1}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ is our **margin of error**
  - **As the confidence level decreases**, z decreases causing the margin of error to decrease, causing the width of the confidence interval to narrow
  - **As the confidence level increases**, z increases causing the margin of error to increase, causing the width of the confidence interval to grow wider

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

Lower Bound

$$(\overline{x_1} - \overline{x_2}) - t_{1-\frac{\alpha}{2}, n_1+n_2-1}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Upper Bound

$$(\overline{x_1} - \overline{x_2}) + t_{1-\frac{\alpha}{2}, n_1+n_2-1}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

We are --% confident that the true population difference of means is between the **lower** and **upper** bound.

Confidence Intervals
Case Two With unknown $\sigma_1 = \sigma_2$

- If all the values on the interval are negative then $\mu_1 < \mu_2$

- If all the values on the interval are positive then $\mu_1 > \mu_2$

- If 0 is on the interval then it's possible that $\mu_1 = \mu_2$

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- We use our sample means to make inference on the population mean

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- $\overline{x_1} - \overline{x_2}$ is our **point-estimate** for the population mean

- $t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is our **margin of error**

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- We use our sample means to make inference on the population mean

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- $\overline{x_1} - \overline{x_2}$ is our **point-estimate** for the population mean

- $t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is our **margin of error**

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- $t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ is our **margin of error**
  - **As either n increases**, $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ decreases, causing the margin of error to decrease causing the width of the confidence interval to narrow
  - **As either n decreases**, $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ increases, causing the margin of error to increase causing the width of the confidence interval to widen
  - **If one decreases and the other increases** – we have to plug in the values to see what the overall effect is

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- $t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ is our **margin of error**
  - **As the confidence level decreases**, z decreases causing the margin of error to decrease, causing the width of the confidence interval to narrow
  - **As the confidence level increases**, z increases causing the margin of error to increase, causing the width of the confidence interval to grow wider

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

Lower Bound

$$(\overline{x_1} - \overline{x_2}) - t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Upper Bound

$$(\overline{x_1} - \overline{x_2}) + t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We are --% confident that the true population difference of means is between the **lower** and **upper** bound.

Confidence Intervals
Case Three With unknown $\sigma_1 \neq \sigma_2$

- If all the values on the interval are negative then $\mu_1 < \mu_2$

- If all the values on the interval are positive then $\mu_1 > \mu_2$

- If 0 is on the interval then it's possible that $\mu_1 = \mu_2$

Example

- According to a NY Times article a survey conducted showed that 22 men averaged about 3 hours of housework per day with a standard deviation of .85 and 49 women averaged about 6 hours of housework per day with a standard deviation of 1.3
- Find a 90% confidence interval for the true population difference of means.

Example

First we solve for v:

$$v = \frac{\left(\frac{.85^2}{22} + \frac{1.3^2}{49}\right)^2}{\frac{\left(\frac{.85^2}{22}\right)^2}{22-1} + \frac{\left(\frac{1.3^2}{49}\right)^2}{49-1}} = 59.5402 \approx 59$$

Example

- We use our sample means to make inference on the population mean

$$(\overline{x_1} - \overline{x_2}) \pm t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(3 - 6) \pm t_{1-\frac{.1}{2},59}\sqrt{\frac{.85^2}{22} + \frac{1.3^2}{49}}$$

$$(-3) \pm (1.671093)\sqrt{\frac{.85^2}{22} + \frac{1.3^2}{49}}$$

$$= (-3.433618, -2.566382)$$

Example

$$(-3.433618, -2.566382)$$

All the values on the interval are negative. This indicates $\mu_1 < \mu_2$ – that the population mean of hours spent doing housework per day for women is higher than it is for males.

Summary!

Sampling Distribution for the Sample Mean Summary

Shape, Center and Spread of Population	Shape of sample	Center of sample	Spread of sample
Populations are normal with means μ and standard deviations σ.	Regardless of the sample size n, the shape of the distribution of the sample mean is normal	$\mu_d = \mu_1 - \mu_2$	$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Population are not normal with means μ and standard deviations σ.	As the sample size n increases, the distribution of the sample mean becomes approximately normal	$\mu_d = \mu_1 - \mu_2$	$s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Confidence Intervals with Paired Data known $\sigma_1 \& \sigma_2$

Assumptions	Point Estimate	Margin of Error
1. Random Sample *2.* $n > 30$ OR the population is bell shaped	$\bar{x}_1 - \bar{x}_2$	$z_{1-\frac{\alpha}{2}}\sqrt{\frac{s_d^2}{n}}$

- We are --% confident that the true difference of population means lays on the confidence interval.

Confidence Intervals unknown $\sigma_1 = \sigma_2$

Assumptions	Point Estimate	Margin of Error
1. *Random Sample* 2. $n > 30$ OR the population is bell shaped	$\bar{x}_1 - \bar{x}_2$	$t_{1-\frac{\alpha}{2}, n_1+n_2-1} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ Where: $s_p^2 = \frac{(\mathrm{n}_1 - 1)\mathrm{s}_1^2 + (\mathrm{n}_2 - 1)\mathrm{s}_2^2}{n_1 + n_2 - 2}$

- We are --% confident that the true difference of population means lays on the confidence interval.

Confidence Intervals unknown $\sigma_1 \neq \sigma_2$

Assumptions	Point Estimate	Margin of Error
1. Random Sample *2.* $n > 30$ OR the population is bell shaped	$\bar{x}_1 - \bar{x}_2$	$t_{1-\frac{\alpha}{2},v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$

- We are --% confident that the true difference of population means lays on the confidence interval.