Lecture 22

Confidence Intervals

STAT 330 - Iowa State University

Confidence Intervals

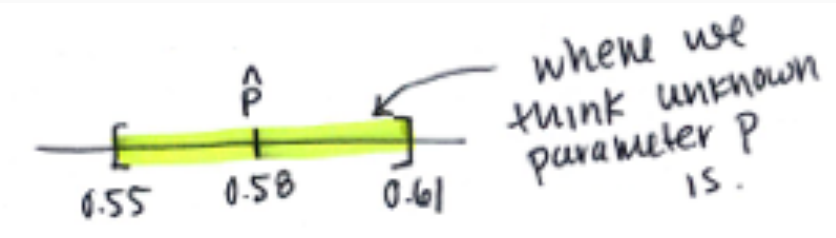
Confidence Intervals

- MLE gives us a "point estimate" of the unknown parameter.
- But $\hat{\theta}$ probably won't *exactly* equal θ due to sampling error.
  $\rightarrow P(\theta = \hat{\theta}) = 0$
- Create a confidence interval to give range of reasonable values for the unknown parameter θ.

Example 1: Polling

Today's poll shows 58% of people favor the new bill. The margin of error is $\pm 3\%$.

The confidence interval for the proportion of people that favor the bill is $[0.55, 0.61]$.



Confidence Interval

Definition

A random interval $[a, b]$ is a $(1-\alpha)100\%$ *confidence interval* for the parameter θ if it contains θ with probability $(1-\alpha)$

$$P(a \leq \theta \leq b) = 1 - \alpha$$

- $(1-\alpha)$ is called the confidence level
- When you estimate an unknown parameter θ, it should be accompanied by a confidence interval

- Interpretation: We are $[(1-\alpha)\%]$ confident that the [insert population parameter + context] is between [insert interval + units].

Constructing Confidence Intervals

In this class, we will construct normal distribution based intervals.

Suppose we have an estimator $\hat{\theta}$ for unknown parameter θ.

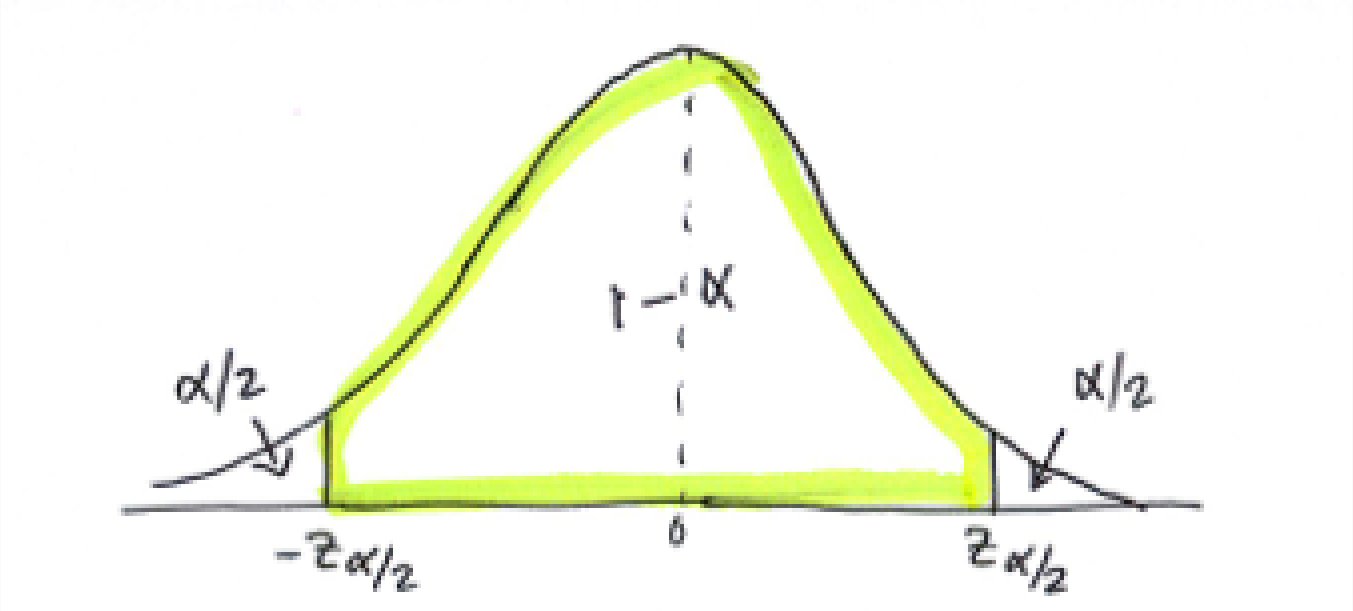
1. $\hat{\theta}$ is unbiased: $E(\hat{\theta}) = \theta$
2. $\hat{\theta}$ follows a normal distribution.

We can standardize $\hat{\theta}$ to get

$$Z = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \sim N(0, 1)$$

where $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})} =$ standard deviation of $\hat{\theta}$

Constructing Confidence Intervals



Let $z_{\alpha/2}$ be the $1 - \frac{\alpha}{2}$ quantile of the standard normal distribution.

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \leq z_{\alpha/2} \right) = 1 - \alpha$$

Constructing Confidence Intervals

Isolating θ in the middle, we get

$$P\left(\hat{\theta} - z_{\alpha/2}SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{\alpha/2}SE(\hat{\theta})\right) = 1 - \alpha$$

Thus, a $(1-\alpha)100\%$ confidence interval for θ is

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

Common choices for α are $0.01, 0.05,$ and 0.1

$(1-\alpha)100\%$	80	90	95	98	99
$z_{\alpha/2}$	1.282	1.645	1.96	2.326	2.576

Constructing Confidence Intervals Cont.

We will make confidence intervals for four cases:

1. μ (population mean)
2. p (population proportion)
3. $\mu_1 - \mu_2$ (difference in population means)
4. $p_1 - p_2$ (difference in population proportions)

Confidence intervals for all 4 of the above cases can be constructed using normal distribution based inference.

All have the general form: $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$. Change the estimator $\hat{\theta}$ based on your problem.

Follow the same general procedure to construct these intervals.

Confidence Interval for Mean

Confidence Interval for μ

Confidence interval for the population mean

$X_1, \ldots, X_n \overset{iid}{\sim} f_X(x)$ with $E(X_i) = \mu$ and $Var(X_i) = \sigma$

First, we estimate μ using the *statistic* $\bar{X}$. From CLT, we know

- $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
- $SE(\bar{X}) = \sqrt{Var(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

A $(1-\alpha)100\%$ confidence interval for μ is

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

In most cases, the population standard deviation σ will be unknown. Replace σ with the sample standard deviation s.

$$\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Confidence Interval for μ Cont.

If we want a 95% confidence interval, then

$$1 - \alpha = 0.95$$
$$\rightarrow \alpha = 0.05$$
$$\rightarrow \alpha/2 = 0.025$$

$z_{\alpha/2} = z_{0.025}$ is the 0.975^{th} quantile of the $N(0, 1)$ distribution.
$\rightarrow$ Using the $z - table$, we get $z_{0.025} = 1.96$.
(Or use the table on slide 6 to obtain $z_{\alpha/2}$)

The 95% confidence interval for μ is

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}} \qquad \text{when } \sigma \text{ is known}$$
$$\bar{X} \pm 1.96\frac{s}{\sqrt{n}} \qquad \text{when } \sigma \text{ is unknown}$$

Example

Example 2: A random sample of 50 batteries were taken for a particular brand. For the sample, the mean lifetime is 72.5 hours and variance is 19.3 hours^2. Find a 95% confidence interval for the true mean lifetime of batteries from that particular brand.

Given

$n = 50$

$\bar{x} = 72.5$

$s^2 = 19.3$

$\rightarrow s = \sqrt{19.3} = 4.39$

95% confidence

$\rightarrow z_{\alpha/2} = 1.96$

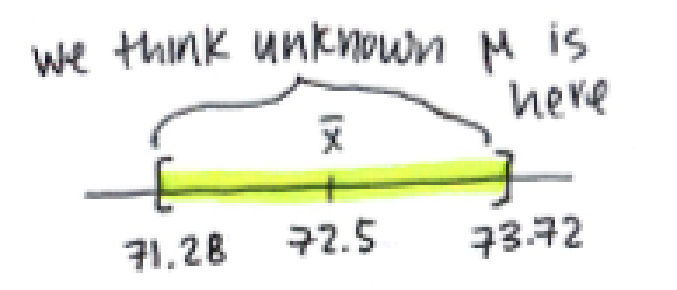
Calculate Confidence Interval

$$\bar{x} \pm z_{/2} SE(\bar{X})$$

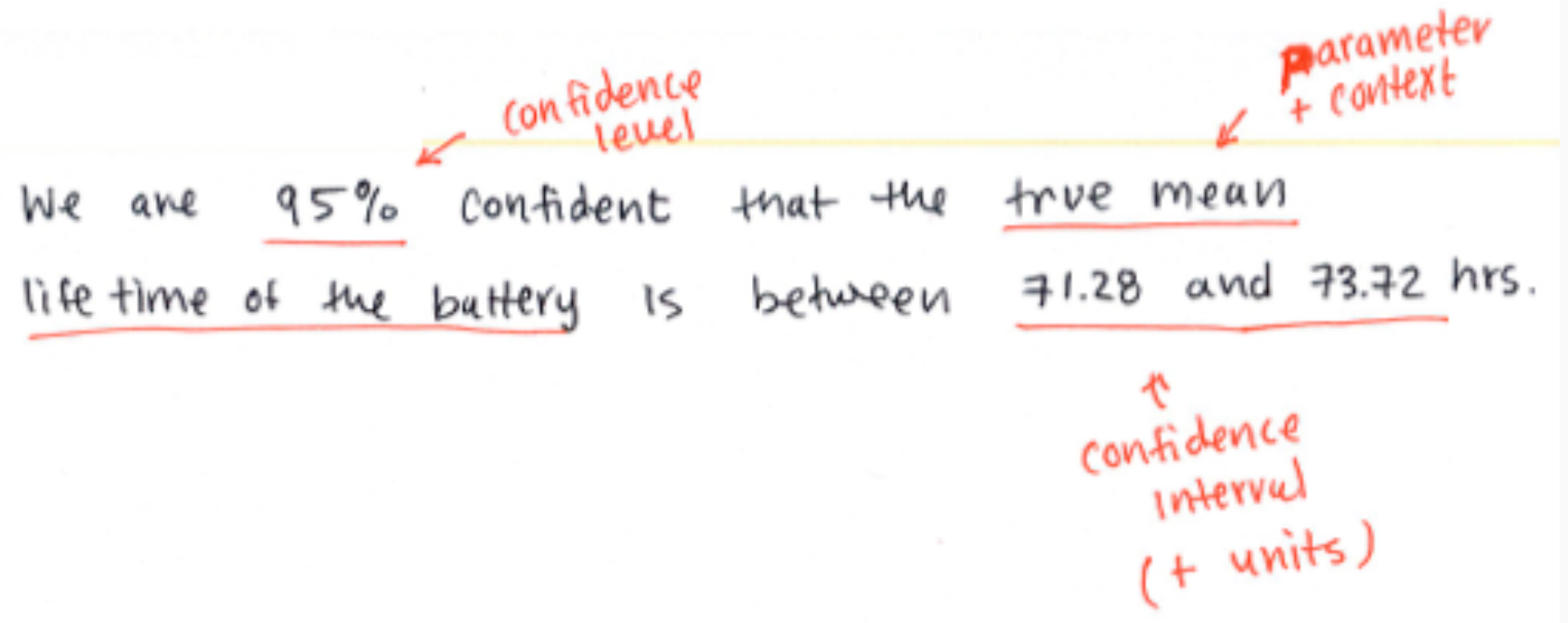
$$\bar{x} \pm z_{/2} \left(\frac{s}{\sqrt{n}} \right)$$

$$72.5 \pm 1.96 \left(\frac{4.39}{\sqrt{50}} \right)$$

$$72.5 \pm 1.2168 \rightarrow (71.28, 73.72)$$



Interpret the Confidence Interval



Confidence Interval for Proportion

Confidence Interval for p

Confidence interval for the population proportion

- In this scenario, we want to estimate the proportion of population belonging to a particular category.
- Any individual in the population either belongs to the category of interest ("1"), or they don't ("0").
- Thus, we can think of each random variable X as a Bernoulli distribution with unknown parameter p
- We ultimately want to estimate and find a confidence interval for p.

  $p = P(Success) = P(\text{being in category of interest})$

Confidence Interval for p Cont.

$X_1, \ldots, X_n \overset{iid}{\sim} Bern(p)$

First, estimate p using the *statistic* $\hat{p} = \frac{\sum X_i}{n}$ = sample proportion.

- $E(\hat{p}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} np = p$ (unbiased)
- $Var(\hat{p}) = Var\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} Var\left(\sum X_i\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

  $\rightarrow SE(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$

Since $\hat{p}$ is the mean of the Bernoulli $X's$, CLT for means applies

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Thus a $(1-\alpha)100\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example

Example 3: In a random sample of 1000 U.S. adults, 38.8% stated they believed in the existence of ghosts. Find a 90% confidence interval for the population proportion of all U.S. adults who believe in the existence of ghosts.

Given

$n = 1000$

$\hat{p} = 0.388$

90% confidence

$\rightarrow z_{\alpha/2} = 1.645$

Calculate Confidence Interval

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.388 \pm 1.645\sqrt{\frac{0.388(1-0.388)}{1000}}$$

$$0.388 \pm 0.0253 \rightarrow (0.363, 0.413)$$

Interpretation: We are 90% confident that the population proportion of all US adults that believe in ghosts is between 0.363 and 0.413.

Confidence Interval for Difference Between Groups

Confidence Intervals

- We have learned how to build a confidence interval to estimate an unknown population parameter θ

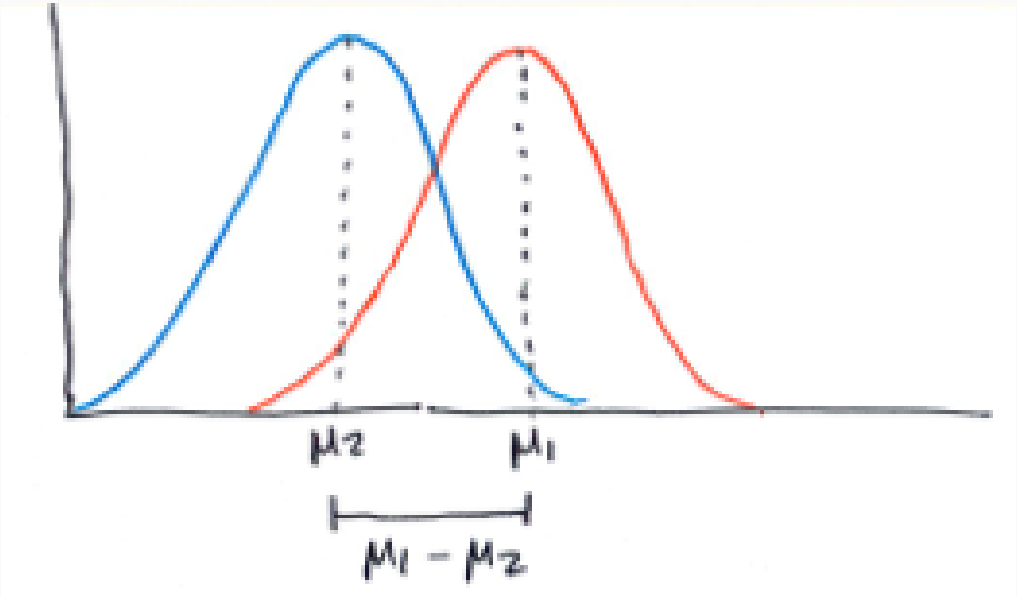
$$\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$$

- Now, we learn how to build a confidence interval to estimate the *difference* between 2 population parameters
  - $\rightarrow$ Compare group 1 and group 2 with parameters θ_1 and θ_2 respectively
  - $\rightarrow$ 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.

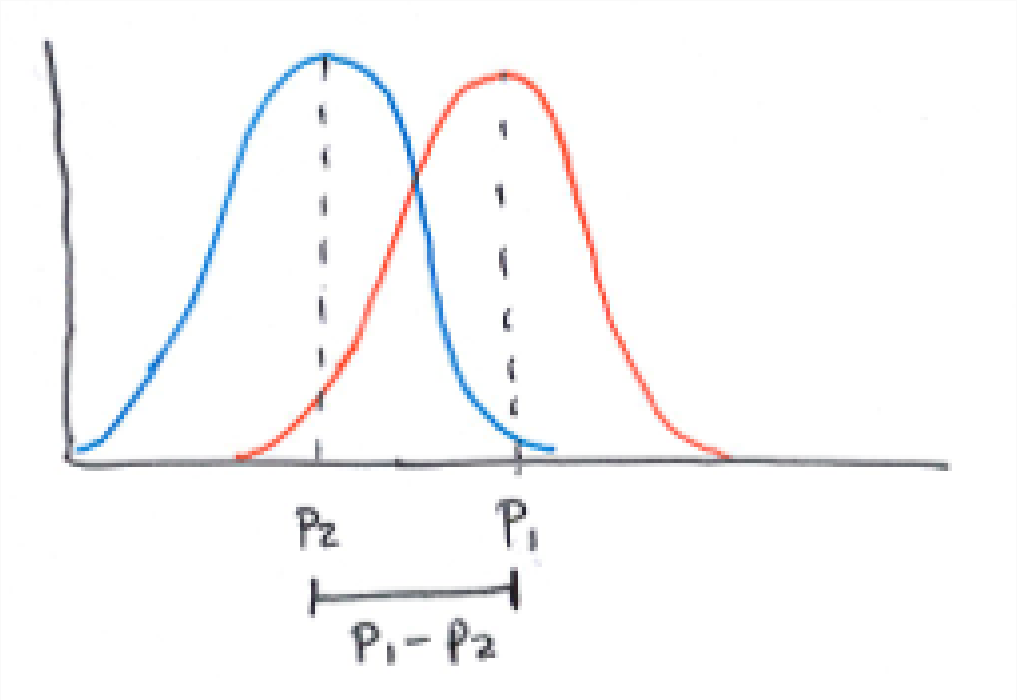
$$\rightarrow SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

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

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

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

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

Examples

Example: Difference in Means

Example 1: Taxable Income

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.

Given

Group 1 (East)	Group 2 (West)
$n_1 = 1000$	$n_2 = 2000$
$\bar{x}_1 = 37200$	$\bar{x}_2 = 42000$
$s_1 = 10100$	$s_2 = 15600$

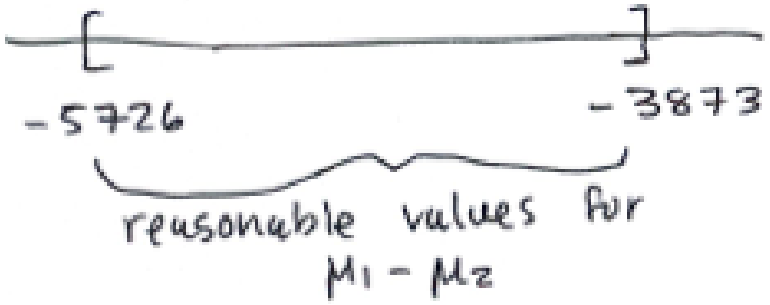
95% Confidence $\rightarrow z_{\alpha/2} = 1.96$

Calculate Confidence Interval

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

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

$$-4800 \pm 927 \rightarrow (-5726, -3873)$$



Since both ends of my CI are negative, we think group 1 − group 2 < 0

⇒ East − West < 0

⇒ 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

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.

Given

Group 1 (A)	Group 2 (B)
$n_1 = 100$	$n_2 = 100$
$\hat{p}_1 = 5/100 = 0.05$	$\hat{p}_2 = 10/100 = 0.10$

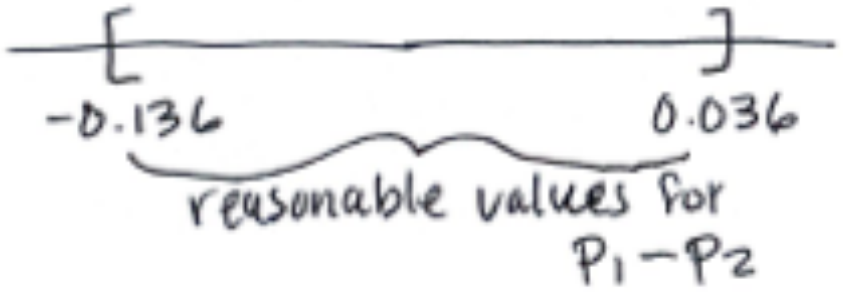
98% Confidence $\rightarrow z_{\alpha/2} = 2.326$

Calculate Confidence Interval

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

$$(0.05 - 0.10) \pm 2.326\sqrt{\frac{0.05(1-0.05)}{100} + \frac{0.10(1-0.10)}{100}}$$

$$-0.05 \pm 0.086 \rightarrow (-0.136, 0.036)$$



lower: -0.136

group1 − group2 < 0

$\Rightarrow A < B$

upper: 0.036

group1 − group > 0

$\Rightarrow A > B$

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.

Confidence Intervals Summary

General formula for confidence intervals: $\hat{\theta} \pm z_{\alpha/2} SE(\hat{\theta})$

Select $z_{\alpha/2}$ value for your confidence level from table:

$(1-\alpha)100\%$ confidence level	80	90	95	98	99
$z_{\alpha/2}$	1.282	1.645	1.96	2.326	2.576

Confidence Interval Formulas:

- CI for mean(μ)

  $\bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

- CI for proportion(p)

  $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

- CI for difference in means $(\mu_1 - \mu_2)$

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

- CI for difference in props$(p_1 - p_2)$

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

Interpretation: We are [C%] confident that [population parameter + context] is between [confidence interval + units].