Hypothesis Testing with P-Values

Administrivia

- **Homework 5** due Friday Nov 10

Previously on CSCI 3022

Def: A **statistical hypothesis** is a claim about the value of a parameter of a population characteristic.

The objective of **hypothesis testing** is to decide, based on sampled data, ***if the alternative hypothesis is actually supported by the data***.

Def: A **test statistic** is a quantity derived from the sample data and calculated assuming that the Null hypothesis is true. It is used in the decision about whether or not to reject the Null hypothesis.

Def: The **rejection region** is a range of values of the test statistic that would lead you to **reject** the Null hypothesis.

We've looked at ways to compute confidence intervals for several different statistics:
E.g. a $100(1-\alpha)\%$ **confidence interval** for the mean μ with known sd. σ is given by

$$\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Critical Region HT Refresher

Example: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

Question: What are the Null hypothesis and alternative hypothesis to test the claim that there is statistical evidence that 1999 Jettas made in Mexico have a smaller life expectancy that those made in Germany.

$H_0: \mu = 300$

$H_1: \mu < 300$

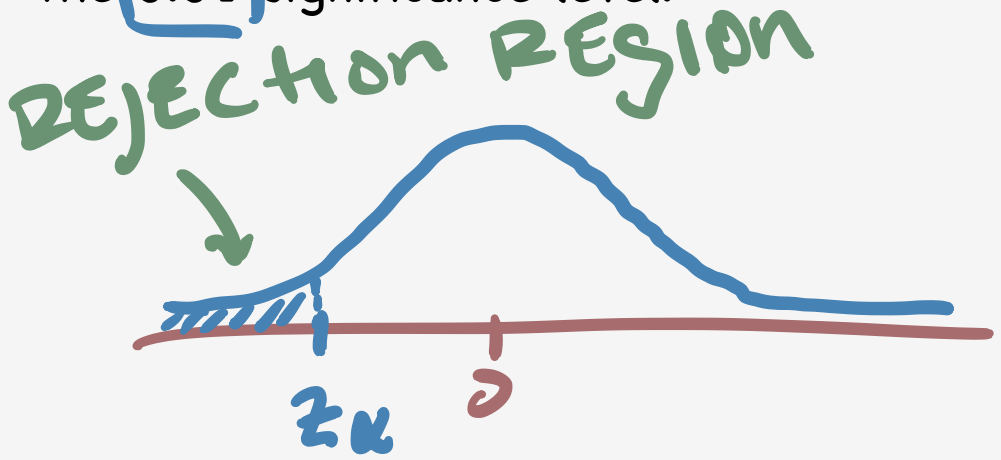
Critical Region HT Refresher

$H_0: \mu = 300$

$H_1: \mu < 300$

Example: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a Critical Region test at the 0.01 significance level.



$\mu = 300, \sigma = 150$

$\alpha = .01 \Rightarrow z_\alpha = -2.33$

$\frac{\bar{x} - \mu}{SE} \sim N(0,1)$

$\rightarrow$ STATS.NORM.PPF(0.99) = 2.33

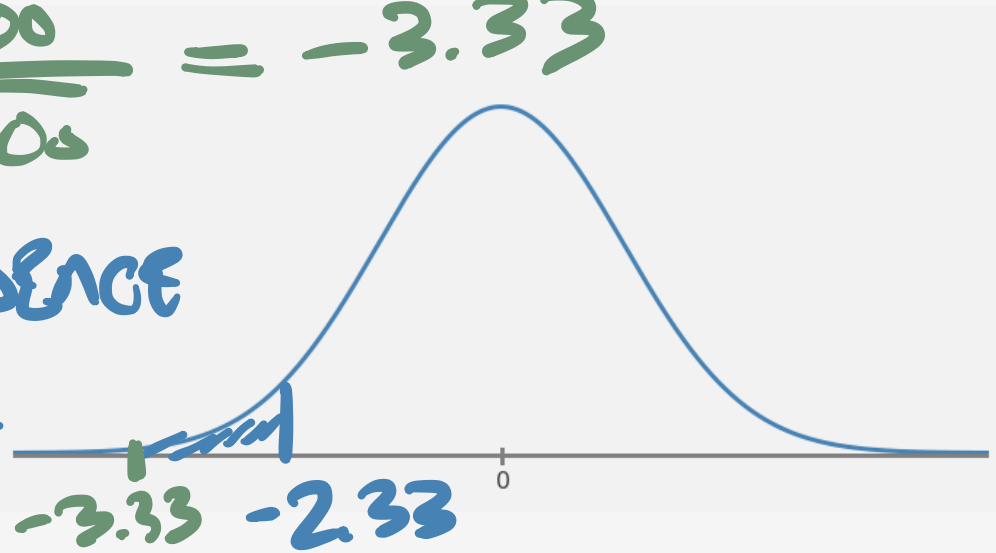
Critical Region HT Refresher

Example: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation ~~100~~ 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be ~~200~~ 250 thousand miles.

Is there sufficient evidence to conclude that, in fact, 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out a Critical Region test at the 0.01 significance level.

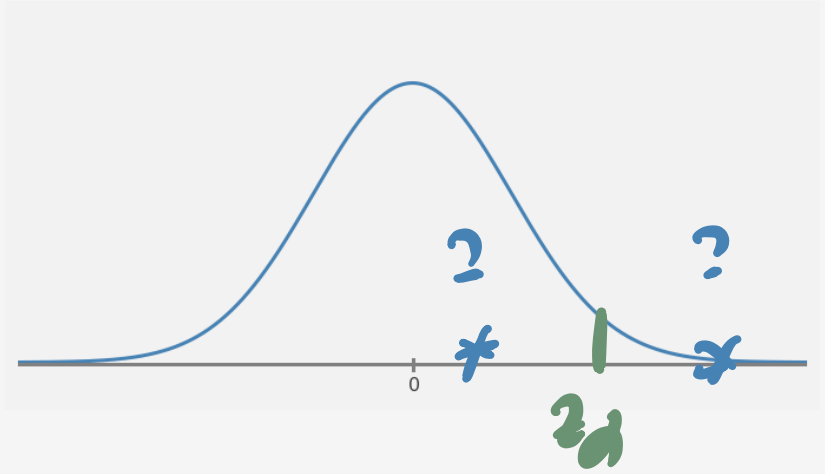
$$\frac{\bar{X}-\mu}{SE} \sim N(0,1) \Rightarrow \frac{250-300}{150/\sqrt{100}} = -3.33$$

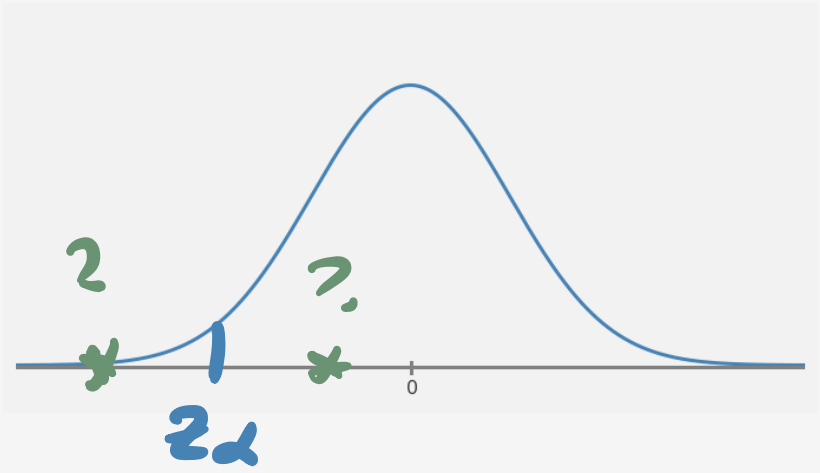
SUFFICIENT STATISTICAL EVIDENCE
⇒ REJECT THE NULL HYP-

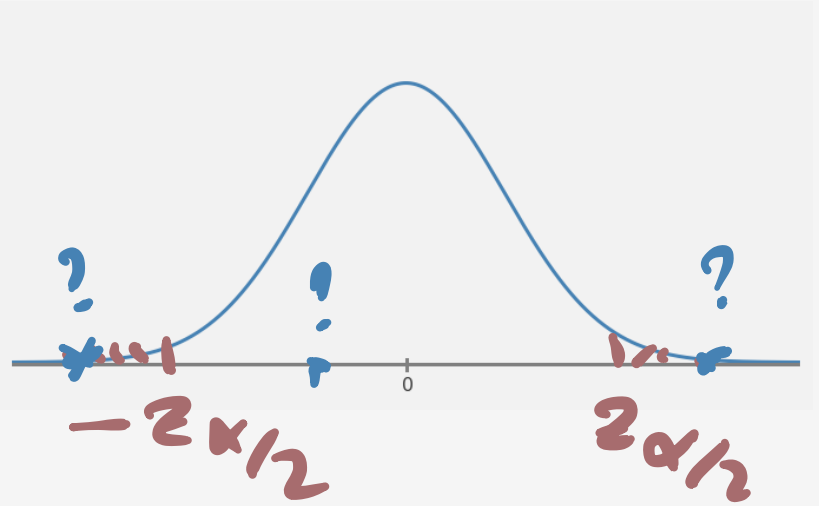


Critical Region HT Summary

Alternative Hypothesis	Rejection Region for Level α Test
$H_1: \quad \theta > \theta_0$	$z \geq z_\alpha$
$H_1: \quad \theta < \theta_0$	$z \leq -z_\alpha$
$H_1: \quad \theta \neq \theta_0$	$(z \leq -z_{\alpha/2})$ or $(z \geq z_{\alpha/2})$







Critical Region HT Summary

- Critical Region is region where test statistic has low probability under Null Hypothesis
- Requires normally distributed data, or large enough sample for Central Limit Theorem
- Under these assumptions we call this a Z-Test
- Rejecting the Null when the Null is true is called a Type I Error
- The probability of committing a Type I Error is α , the significance level of the test
- Failing to reject the Null when the Null is false is called a Type II Error

Introduction to P-Values

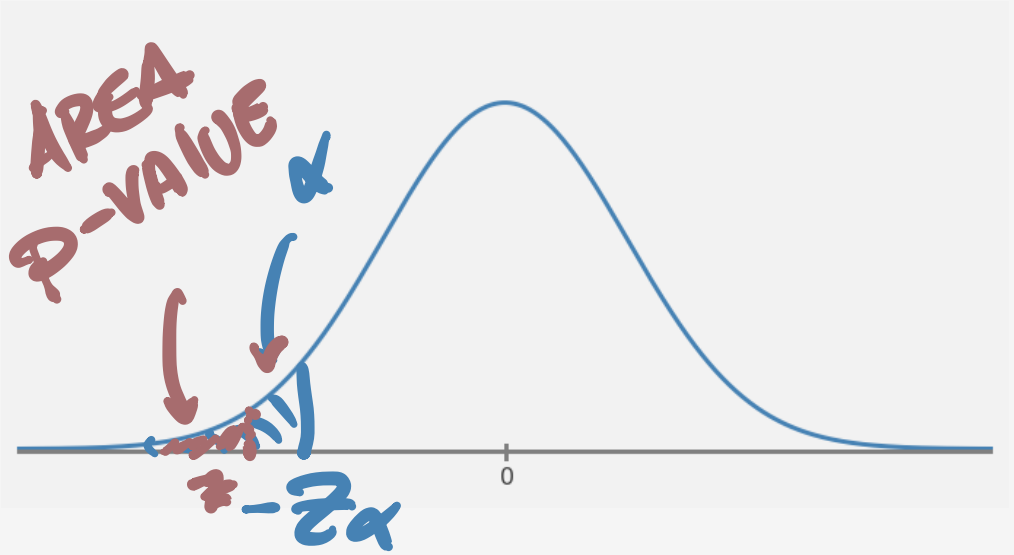
Another way to view the critical region hypothesis test is through a so-called **p-value**

This framework for HT is very popular in scientific study and reporting

Example: Consider a lower-tail critical region test with the following hypotheses:

$$\begin{aligned} H_0: &\quad \mu = \mu_0 \\ H_1: &\quad \mu < \mu_0 \end{aligned}$$

The critical region test is:



Introduction to P-Values

Def: A **p-value** is the probability, under the Null hypothesis, that we would get a **test statistic at least as extreme as the one we calculated**.

Def: For a lower-tailed test with test statistic x, the p-value is equal to $P(X \leq x \mid H_0)$

Intuition: The p-value assesses the extremeness of the test statistic. The smaller the p-value, the more evidence we have against the Null hypothesis

Important Notes:

- The p-value is calculated under the assumption that the Null hypothesis is true
- The p-value is always a value between 0 and 1
- The p-value is **NOT** the probability that the Null is true!!

The P-Value Decision Rule

As before, select a significance level α before performing the hypothesis test

Then the decision rule is:

- If p-value $\leq \alpha$ then reject the Null hypothesis
- If p-value $> \alpha$ then fail to reject the Null hypothesis

Thus if the p-value exceeds the selected significance level then we cannot reject the Null hypothesis

Note: The p-value can be thought of as the smallest significance level at which the Null hypothesis can be rejected.

Jetta Life Expectancy with P-Values

Example: The 1999 Volkswagen Jetta was one of the first VW models produced both in Germany and Mexico. The life expectancy of Jettas produced in Germany was found to follow a normal distribution with mean 300 thousand miles and standard deviation 150 thousand miles. Life expectancy of models made in Mexico were recorded for a sample of size 100. The sample mean of these Jettas was found to be 250 thousand miles.

Is there sufficient evidence to conclude that 1999 Jettas made in Mexico have a shorter life expectancy that those made in Germany? Carry out the p-value test at the 0.01 SL.

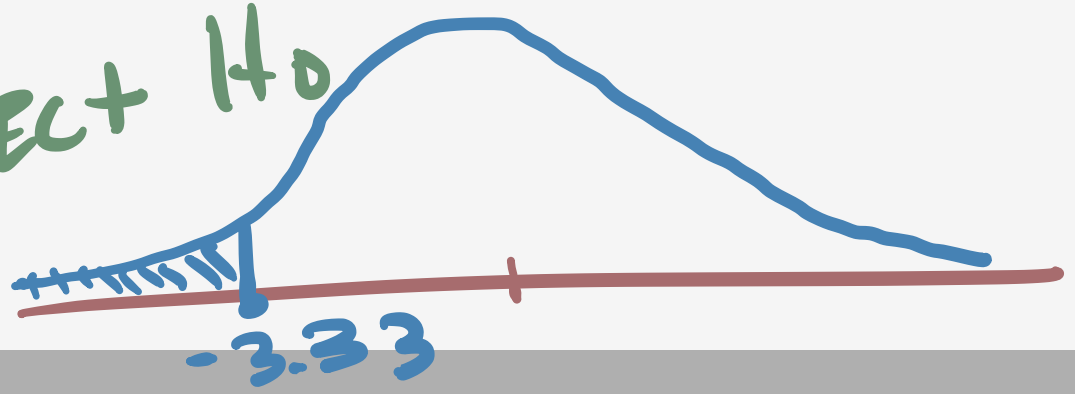
TEST STATISTIC

$$\frac{250 - 300}{150/\sqrt{100}} = -3.33$$

P-VALUE $= \Phi(-3.33)$

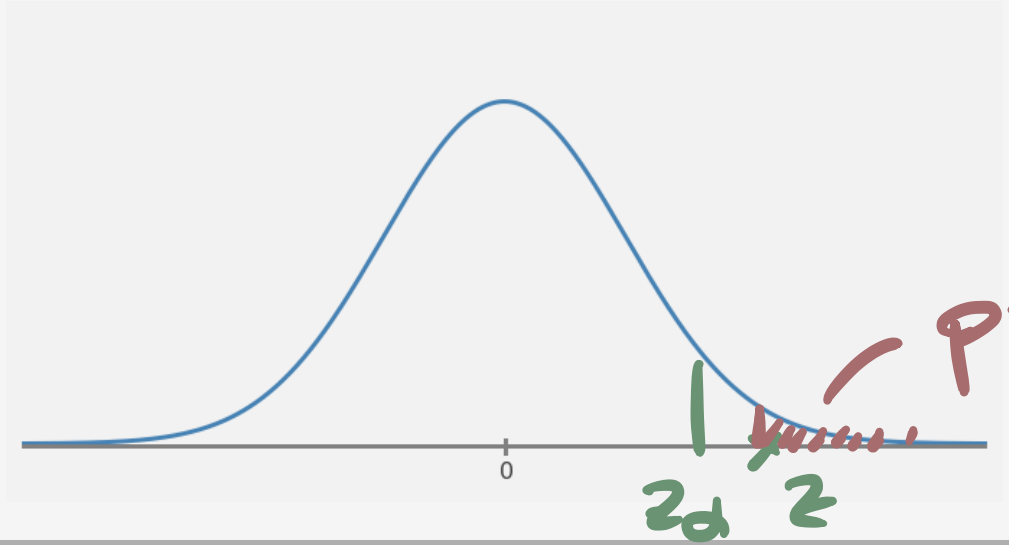
$= 0.00043$

$0.00043 \overset{?}{\leq} 0.01 \Rightarrow$ REJECT H_0



P-Values for Different Z-Tests

Alternative Hypothesis	Critical Region Level α Test	P-Value Level α Test
$H_1: \quad \theta > \theta_0$	$z \geq z_\alpha$	if $1-\Phi(z) \leq \alpha$
$H_1: \quad \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1: \quad \theta \neq \theta_0$	$(z \leq -z_{\alpha/2})$ or $(z \geq z_{\alpha/2})$	



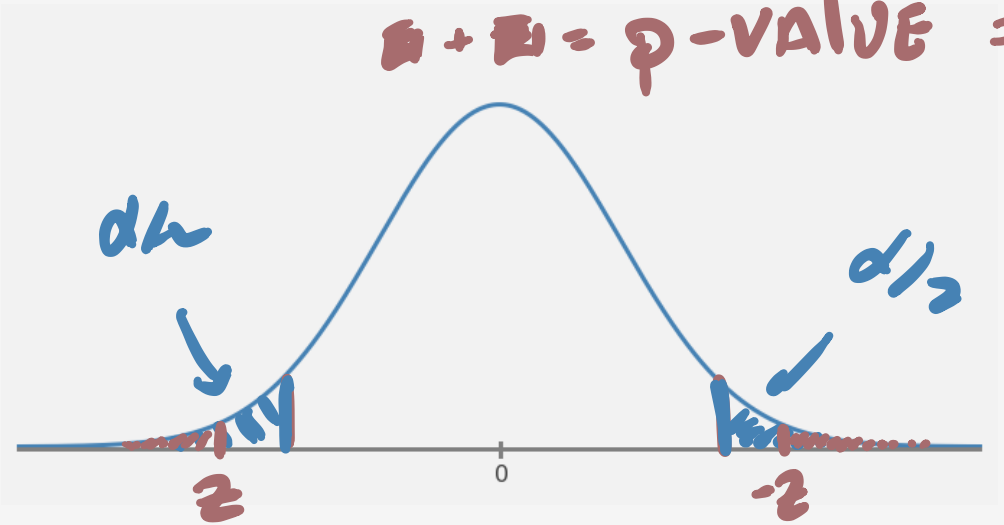
P-Values for Different Z-Tests

Alternative Hypothesis	Critical Region Level α Test	P-Value Level α Test
$H_1: \quad \theta > \theta_0$	$z \geq z_\alpha$	$\Phi(z) \leq \alpha$
$H_1: \quad \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1: \quad \theta \neq \theta_0$	$(z \leq -z_{\alpha/2})$ or $(z \geq z_{\alpha/2})$	

P-Values for Different Z-Tests

Alternative Hypothesis	Critical Region Level α Test	P-Value Level α Test
$H_1: \quad \theta > \theta_0$	$z \geq z_\alpha$	
$H_1: \quad \theta < \theta_0$	$z \leq -z_\alpha$	
$H_1: \quad \theta \neq \theta_0$	$(z \leq -z_{\alpha/2})$ or $(z \geq z_{\alpha/2})$	

$\min\{\Phi(z), 1-\Phi(z)\} \times 2$



α/2 + α/2 = p-value = $2 \times \Phi(-|z|)$

Is the Belgian 1 Euro Biased?

Example: To test if the Belgian 1 Euro coin is fair you flip it 100 times and observe 38 Heads. Perform a p-value Z-test at the .05 significance level.

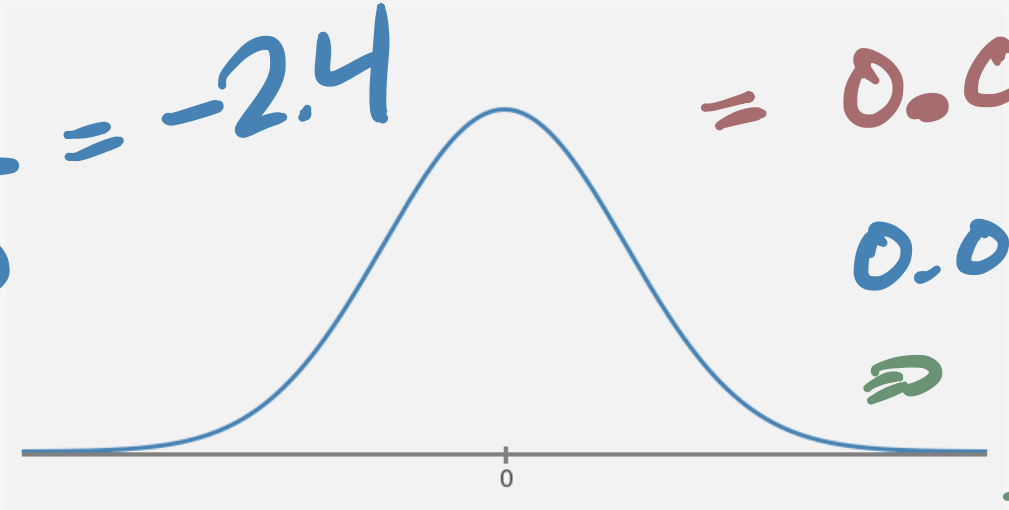
$\alpha = .05$

$H_0 : p = .5$

$H_1 : p \neq .5$

$\hat{p} = 0.38$

$$z = \frac{(.38 - .5)}{\sqrt{\frac{.5(1-.5)}{100}}} = -2.4$$



$2 \times \Phi(-1 - 2.4|)$

$= 2 \times \Phi(-2.4)$

$= 0.0164 = p\text{-value}$

$0.0164 \leq 0.05$

$\Rightarrow$ REJECT H_0

THAT COIN IS FAIR

Is the Belgian 1 Euro Biased?

Example: To test if the Belgian 1 Euro coin is fair you flip it 100 times and observe 38 Heads. Perform a p-value Z-test at the .05 significance level.

Two-Sample Testing for Difference of Means

Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value.

Question: What kinds of Null and alternative hypotheses might we want to test?

$H_0: \mu_1 - \mu_2 = c$

$H_1: \mu_1 - \mu_2 > c$

$H_1: \mu_1 - \mu_2 < c$

$H_1: \mu_1 - \mu_2 \neq c$

$$\frac{(\mu_1 - \mu_2) - c}{\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}} \approx N(0,1)$$

Two-Sample Testing for Difference of Means

Suppose we want to test whether or not the difference of sample means from two populations is nonzero, or equal to a particular value.

Assuming that our sample sizes are large enough, we can standardize our test statistics as:

We can then compute an appropriate p-value in the usual way

Two-Sample Testing for Difference of Means

Example: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
No	663	μ_2 2258	1519 s_2
Yes	413	μ_1 2637	1138 s_1

Does this data provide statistical evidence at the 0.05 significance level that true average calorie intake for teens who typically eat fast food exceeds that of teens who do not typically eat fast food by more than 200 calories per day?

$H_0: \mu_1 - \mu_2 = 200$

$H_1: \mu_1 - \mu_2 > 200$

$$z = \frac{(2637 - 2258) - 200}{\sqrt{\frac{1138}{413} + \frac{1519}{663}}} = 2.20$$

p-value $= 1 - \Phi(2.20) = 0.014 < .05 \Rightarrow$ Reject H_0

Two-Sample Testing for Difference of Means

Example: Data on calorie intake both for a sample of teens that reported that they do not typically eat fast food and another sample of teens who said they did usually eat fast food is as follows:

Fast Food	Sample Size	Sample Mean	Sample SD
No	663	2258	1519
Yes	413	2637	1138

Does this data provide statistical evidence at the 0.05 significance level that true average calorie intake for teens who typically eat fast food exceeds that of teens who do not typically eat fast food by more than 200 calories per day?

Common P-Value Misunderstandings

Misconception #1: If p = 0.05, the Null hypothesis only has a 5% chance of being true.

Common P-Value Misunderstandings

Misconception #2: If p is very small then your alt hypothesis is very likely to be significant

Common P-Value Misunderstandings

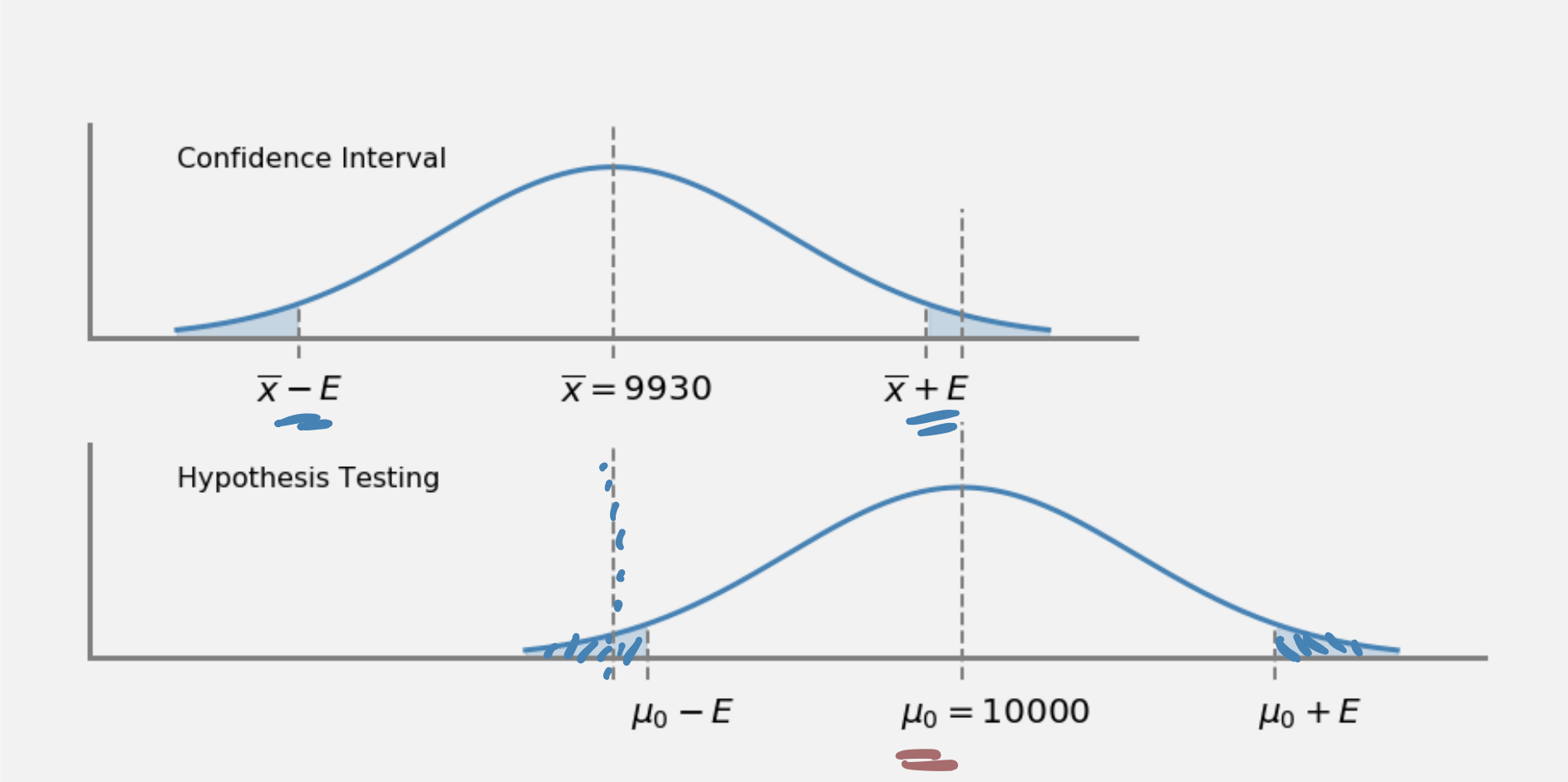
Misconception #3: A statistically significant effect is equivalent to a substantial effect

CIs vs Critical Regions vs P-Values

Confidence Intervals, Critical Regions, and P-Values are three sides to the same coin

CIs vs Critical Regions vs P-Values

Confidence Intervals, Critical Regions, and P-Values are three sides to the same coin



OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 18 In-Class Notebook

Let's:

- Work through some more hypothesis test examples using p-values