Week 2
Hypothesis Testing & Experiments

Applied Statistical Analysis/Quantitative Methods I

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2024

Roadmap for today

Last class we learned:

- What's a sampling distribution (e.g., $\bar{y}$)
- How to create CI around point estimate of a parameter

Roadmap for today

This class:

- What is a hypothesis test?
  - We will test hypotheses about population from sample
- Relationship between CI and null hypothesis testing
- Five steps of hypothesis testing
- Types of errors
- Discussion of 1-sided/2-sided tests

Sometimes, our specific aim is to understand if X causes Y

- Compare two independent samples (`t.test()`)

Hypothesis testing: The big picture

The goal of hypothesis testing is to see if the data agree with some prediction we make based on our theory

> ***Definition 1***: *A hypothesis is just some statement about a characteristic of a variable or a set of variables*

Hypothesis testing: The big picture

***Definition 2**: In statistics, a hypothesis is a statement about a population*

- *It is a prediction: We believe a parameter describing a variable takes a particular numerical value or falls in a certain range of values*

To test a hypothesis, we take our data and conduct a *significance test* (i.e. how likely is it that data support my hypothesis?)

CIs vs. hypothesis testing

- Confidence intervals are about estimating a parameter value
- Hypothesis testing is about testing scientific claims about the value of a parameter
- Procedure is general, but can test a single population mean or compare 2 means

Can't stress this enough: **Everything we do for the rest of this class is basically hypothesis testing in one form or another!**

Step 1/5: Assumptions about your data

- We have to make some assumptions about the data and where it came from
- These assumptions determine how (and if) we can test our hypothesis

Important assumptions include:

- Type of data (continuous, categorical, etc.)
- Sample size
- Sampling method (i.e., randomization)

Step 2/5: Null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction* (disprove the null)

(1) Set up a null hypothesis (H_o)

(2) Use a sample statistic to show that data we have observed would be very unlikely if null hypothesis were true

(3) Derive conclusion about if our H_o is true (kind of)

- This is usually called the alternative hypothesis (H_a)

To do this, we first need to specify our hypotheses!

Step 3/5: Calculate a Test Statistic

We calculate a statistic that summarizes how much our data differs from what we would have expected to observe *if the null hypothesis* were **true**

- Usually this is something equivalent to a Z- or t-statistic

Step 4/5: Calculate a P-Value

*A **P-Value** is a measure of surprise*

- *We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- Smaller P-Values more strongly contradict the null
- Might not be intuitive right away, that's okay P-Values are very difficult for many people to understand!

Step 5/5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ "statistically significant" result
- We would observe a test-statistic this extreme or more 1/20 times if H_0 was true
- More generically, we want $p \leq \alpha$
- We “reject the null” and conclude that evidence supports H_a

Large sample significance testing for means

- Step 1: Assumptions
  - $n \geq 30$
  - Quantitative data (i.e., not a proportion but we can do this)
- Step 2: State hypotheses
  - $H_0 : \mu = \mu_0$
  - This is a "2-sided test," but it may be a "1-sided"

Step 3: Calculate a test statistic

- $TS = \dfrac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}$
- TS is our "test statistic"
- Just as before, this comes from the sampling distribution of $\bar{Y}$

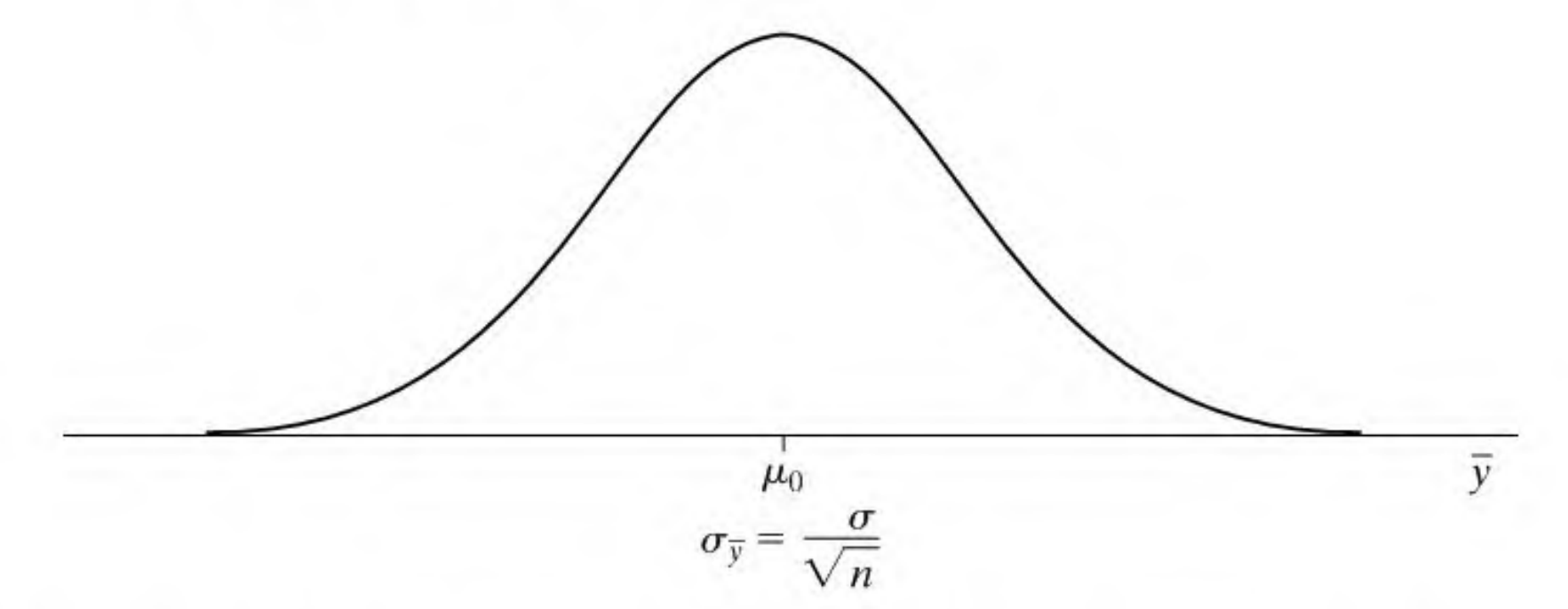
Step 4: P-Value

$$p = Pr(Z \geq |\frac{\bar{Y} - \mu_o}{\sigma_{\bar{Y}}}|) + Pr(Z \leq -|\frac{\bar{Y} - \mu_o}{\sigma_{\bar{Y}}}|) \quad (1)$$

$$= 2 \times Pr(Z \geq |\frac{\bar{Y} - \mu_o}{\sigma_{\bar{Y}}}|) \quad (2)$$

- We use both tails because we want to find the probability of error in both directions (since $H_o : \mu = \mu_o$, we need to check above and below)

Step 4: P-Value



Remember, p-value is referencing sampling distribution of $\bar{y}$

Step 5: Draw a conclusion

- If $p \leq \alpha$ we conclude that the evidence supports H_a
- If $p > \alpha$ we say that "we cannot reject the null hypothesis"

Example: Women "like" the Democratic party

2020 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100)

- We'll just use data from self-identified female respondents
- We are going to assume, for class, that any thermometer rating > 58 means you "like" the Democratic party
- Test hypothesis that women are not neutral towards the Democratic party (they either like or dislike them)
- A score of exactly 58 means the respondent is "neutral"

Example: Women "like" the Democratic party

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample
- What's our null? $H_0 : \mu = 58, H_a : \mu \neq 58$
- Z-score? $= \frac{60.5-58}{1} = 2.5$
- $P - value$ =`2*pnorm(-abs(2.5))`
  $2 \times 0.006 = 0.012$
- Is that good enough?
- Why are we using a two-sided test?

Small sample significance testing

Step 1: Assumptions

- Random sampling
- Quantitative data
- **Population is distributed normally**

Step 2: State hypotheses

- $H_0 : \mu = \mu_0$ (e.g., $\mu = 12$)
- $H_a : \mu \neq \mu_0$
- This is a "2-sided test," but it may be "1-sided". How?

Step 3: Calculate a test statistic

- $t^* = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}, df = (n-1)$
- Just as before, this comes from sampling distribution of $\bar{Y}$
- We'll reference the t-table rather than the z-table, why?

Step 4: P-Value

- $p = Pr\left(|t| \geq |\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}|\right) = 2 \times Pr(t \geq |\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}|)$
- Make sure you are using the right degrees of freedom
- We use both tails because we want to find the probability of error in both directions
- `2*pt(abs(`t^*`), df = n-1, lower.tail=F)`

Step 5: Draw a conclusion

- If $p \leq \alpha$ we conclude that the evidence supports H_a
- But always report the p-value

Example: Country spending on education

Our theory: Countries are not spending 5% of their income on education

The data indicate that:

- $\bar{Y} = 4.7, S = 0.0922$
- TS = $t^* = \frac{4.7-5}{0.09/\sqrt{50}} = -2.279, df = 49$
- P-value = `2*pt(2.279, df=49, lower.tail = F)` = 0.027

Why run a 2-tailed test?

Proportions: Large sample significance testing

Step 1: Assumptions

- Random sampling
- Quantitative data
- $n \geq \frac{10}{min(\pi_0, 1 - \pi_0)}$
  - $min()$ means the minimum of the 2 numbers
  - ≥ 10 is sort of arbitrary (it's a good rule of thumb)
- If our n is bigger than this, we can use the calculations below

Step 2: State hypotheses

- $H_0 : \pi = \pi_0$ (e.g., $\pi_0 = 0.5$)
- $H_a : \pi \neq \pi_0$
- This is a "2-sided test," but it may be a "1-sided". How?

Step 3: Calculate a test statistic

- $Z^* = \dfrac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}}$,

  $\sigma_{\pi_0} = \sqrt{\dfrac{\pi_0(1-\pi_0)}{n}}$

- Just as before, this comes from the sampling distribution $\hat{\pi}$ that would exist if H_0 were true

- Note: we're assuming that $\pi = \pi_0$ to calculate the SE

Step 4: P-Value

- $p = Pr(Z \geq |\frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}}|) + Pr(Z \leq -|\frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}}|) = Pr(|Z| \geq |\frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}}|)$
- $= 2 \times Pr(Z \geq |\frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}}|)$
- We use both tails , why?

Step 5: Draw a conclusion

- If $p \leq \alpha$ we conclude that the evidence supports H_a
- You should **still** report the p-value

Example: Female Board members of Fortune 500

Self-identified women are (roughly) 50% of the U.S. population

- Are they represented (descriptively) in board rooms?
- Use data from 2023 fiscal year, we see that there are $\approx$ 1700 female board members[1]
- What's our null? $H_0 : \pi = 0.5$, $H_a : \pi \neq 0.5$
- $n = 5670, \hat{\pi} = \frac{1700}{5670}$
- Are assumptions met?

[1] Fortune 2023

Example: Female Board members of Fortune 500

Are assumptions met?

- $n \geq \frac{10}{min(.5, 1-.5)} = 20$
- Since our n is bigger, we can proceed
- $Z = \frac{\hat{\pi} - 0.5}{\sigma_{\pi_0}} = \frac{0.3-0.5}{\sqrt{\frac{.5(1-.5)}{5670}}} = \frac{-0.2}{0.007} \approx -28.57$
- `2*pnorm(-28.57, lower.tail=F)`
  - $\approx 0 \Rightarrow$ reject null

Type 1 and Type II Error

Imagine we're all jurors, what if we sentence an innocent man?

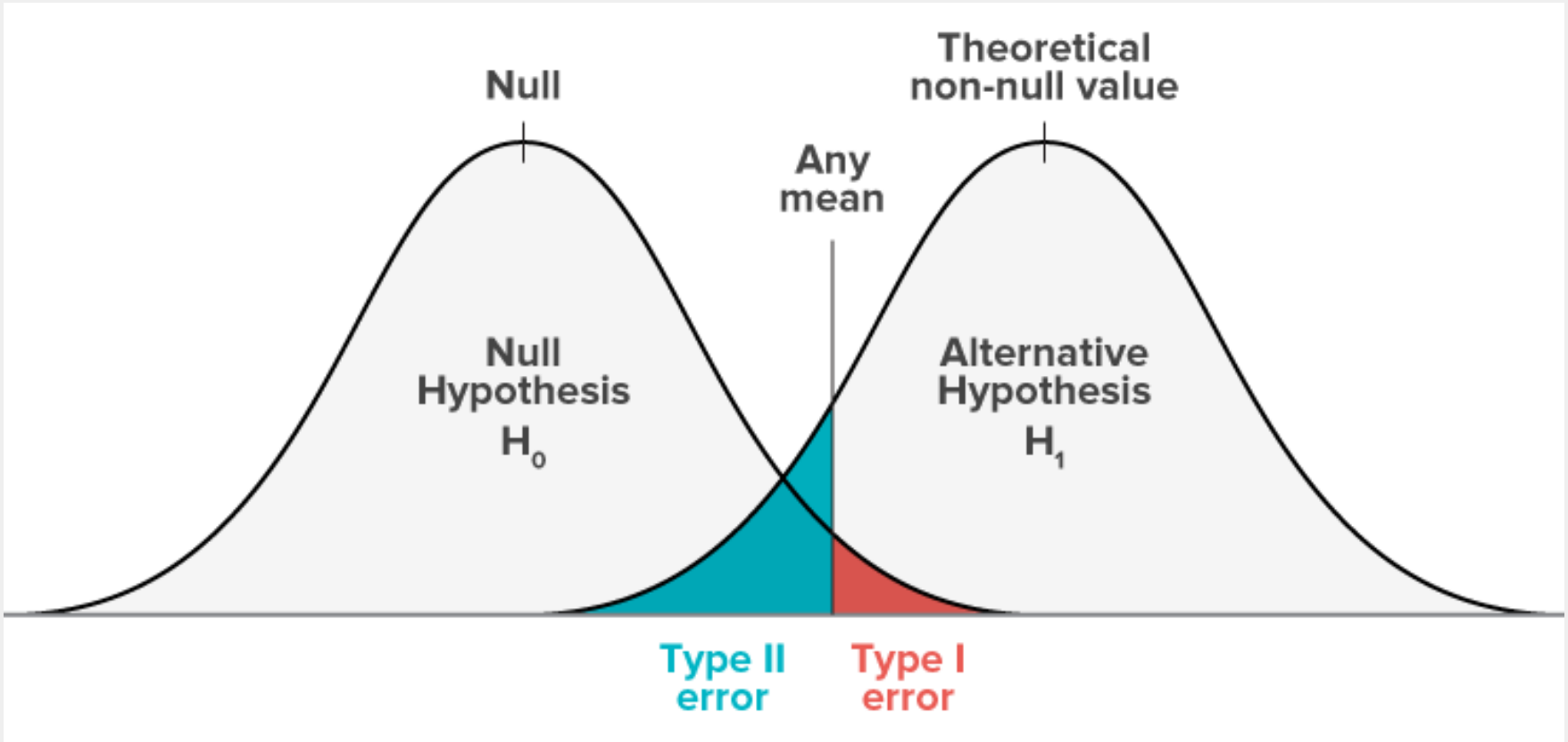
		Jury decision	
		Guilty	Innocent
Truth	Guilty	Correct	Type II
	Innocent	Type 1	Correct

Let's put the null hypothesis on trial!

		Researcher Conclusion	
		Reject Null	Don't reject
Truth	Null is False	Correct	Type II
	Null is True	Type 1	Correct

- Type I error is when we reject a null hypothesis when the null it is actually true
- Type II error is when we fail to reject a null hypothesis when the null is actually false
- We tend to prioritize reducing Type I error, although there are trade-offs

Visualization of Type I and Type II error



How to calculate Type I and Type II error

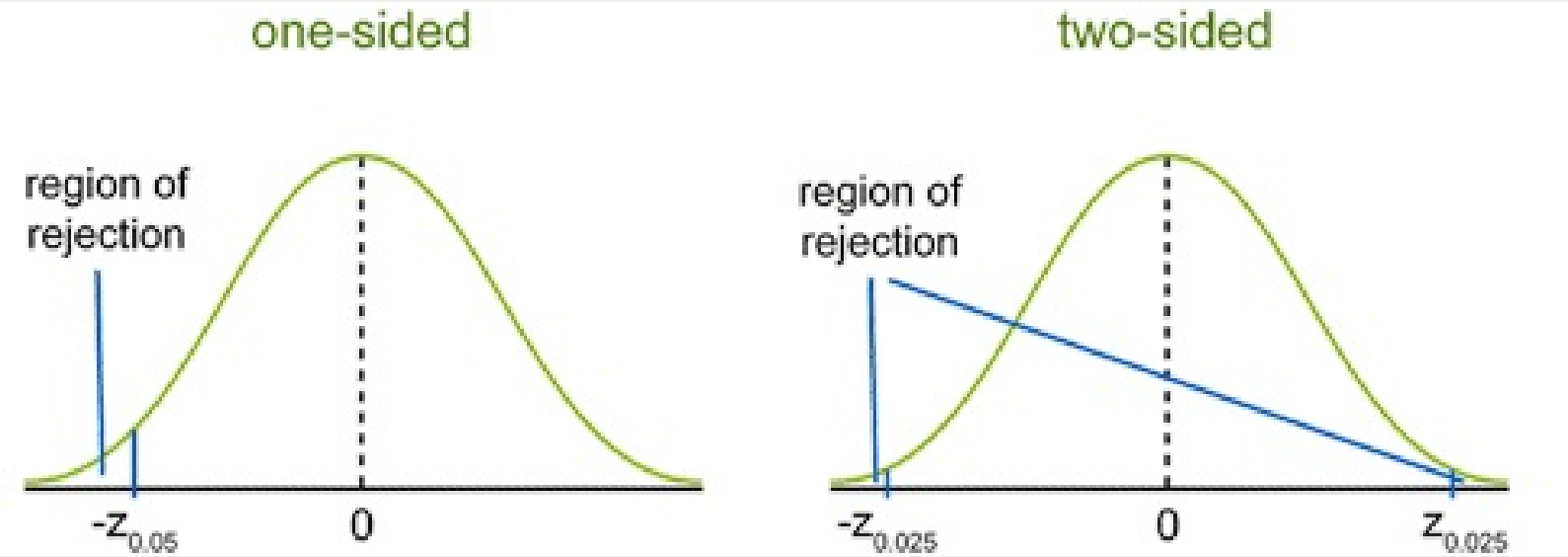
Type I:

- $Pr(Reject\ H_o | H_o True) = \alpha$
- This is our primary focus

Type II (hard):

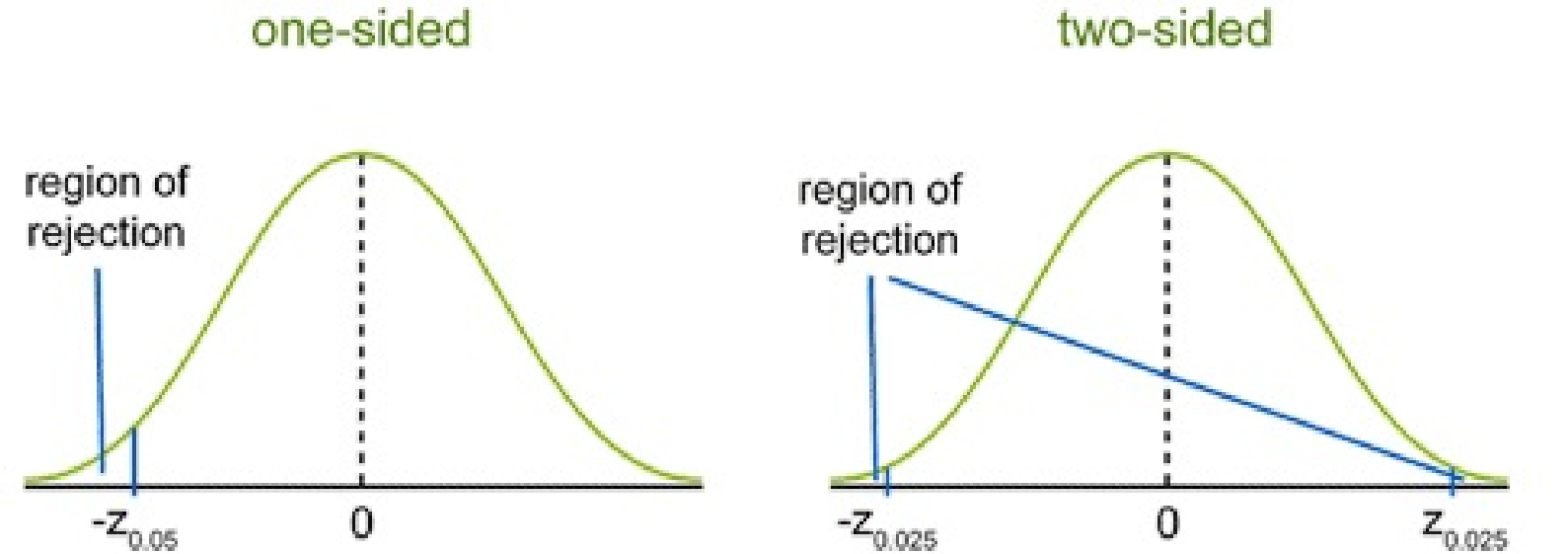
- $Pr(Fail\ to\ reject | H_o\ False)$
- H_o can be false for many values
- Can calculate the strength or "power" of a hypothesis test
  - $1 - Pr(Type\ II\ error)$
  - E.g., probability of finding an effect, assuming that the effect is actually there
- Remember: trade-off between Type I and Type II error

Let's dive deeping into: Critical values



- You can think of hypothesis testing as, "Is the t-statistic so extreme that it is in the rejection region?"
- If $|Z^*|$ is bigger than some "critical value" (Z_α), then p-value will be smaller than α

Critical values: which tail?



- Note that the p-value calculations will be different if you are only calculating the area under 1-tail
- When you calculate the p-value for one-sided tests, make sure you are calculating it for the correct tail

Critical values: 1 sided vs. 2 sided

Central message: Look in the tail associated with H_A

Go back to our example of female board members

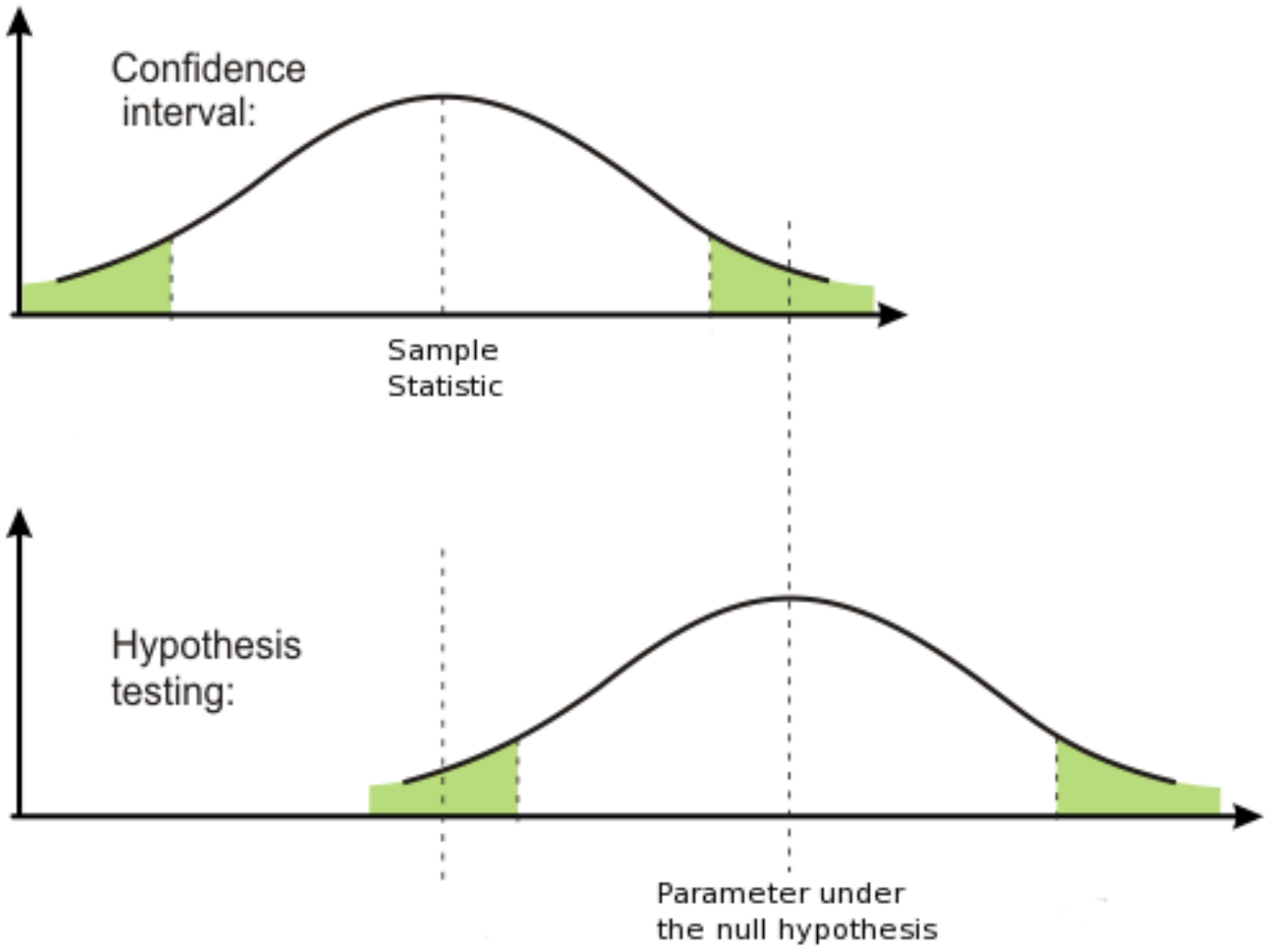
- This time let $H_0 : \pi \geq 0.5, H_a : \pi < 0.5$
- $Z = \frac{\hat{\pi} - 0.5}{\sigma_{\pi_0}} = \frac{0.3 - 0.5}{\sqrt{\frac{.5(1-.5)}{5670}}} = \frac{-0.2}{0.007} \approx -28.57$
- Note: still using $\pi_0 = 0.5$
  - Smallest value possible for π in region covered by H_0
- We use $P(Z \geq |\frac{\hat{\pi} - 0.5}{\sigma_{\pi_0}}|)$ ... not $|Z|$
- `pnorm(-28.57, lower.tail=T)` $< 0.001 \Rightarrow$ reject null

Example: 1 sided vs. 2 sided

Now let $H_0 : \pi \leq 0.5, H_a : \pi > 0.5$

- Still $Z = \frac{\hat{\pi} - 0.5}{\sigma_{\pi_0}} = \frac{0.3 - 0.5}{\sqrt{\frac{.5(1-.5)}{5670}}} = \frac{-0.2}{0.007} \approx -28.57$
- Now, we want $P(Z \leq |\frac{\hat{\pi} - 0.5}{\sigma_{\pi_0}}|)$
- `pnorm(-28.57, lower.tail=F)` $\approx 1 \Rightarrow$ don't reject H_0
- Note: `lower.tail = F`

Confidence intervals & null hypothesis testing



Example: Confidence interval approach #1

According to a union agreement, the true mean income for all senior-level assembly-line workers in a large company equals €525 per week. A representative of a women's group decides to analyze whether the mean income μ for female employees matches this norm. For a random sample of 36 female employees, $\bar{y} = €495$ and $s = €75$.

- Let's use a 95% CI, or $\alpha = .05$, and assume that the CLT applies (no T-distribution)
- $525 \pm 1.96 * \sigma_{\bar{y}} = 525 \pm 1.96 \frac{s}{\sqrt{n}}$
- $= 525 \pm 1.96 \frac{75}{\sqrt{36}}$
- $525 \pm 1.96 \times 24.5$
- 95% CI = $[500.5, 549.5]$

$H_0 : \mu = \bar{y}$, since we observed $\bar{y} = 495$, reject H_0

Example: Confidence interval approach #2

- Let's use a 95% CI, or $\alpha = .05$
- $495 \pm 1.96, \sigma_{\bar{y}} = 495 \pm 1.96 \frac{S}{\sqrt{n}}$
- $= 495 \pm 1.96 \frac{75}{\sqrt{36}}$
- $495 \pm 1.96 \times 24.5$
- 95% CI = $[470.5, 519.5]$

Since $H_0 : \mu = 525$ is not in that interval, we can reject H_0

Now that we can do hypothesis testing...

1. We should think about causality
2. Average treatment effects
3. When we don't have random assignment (and now we're worried about pre-treatment confounders), what do we do?
   - Contingency tables, regression, etc.

First, what is causality?

In social science we want to make causal claims

$$X \rightarrow Y$$

What does this mean?

- Let's use some formal notation using an example experiment

We will use T to represent a **treatment variable**

- For a categorical treatment, which treatment did person i receive?

$$T_i = \begin{array}{l} 1 \text{ if unit } i \text{ receives the treatment} \\ 0 \text{ if unit } i \text{ receives the control} \end{array}$$

- We let y_i^1 represent the outcome of the i^{th} unit **if the treatment is given**
- We let y_i^0 represent the outcome of the i^{th} unit **if the control is given**
- One of these is observed, the other is the **counterfactual** – what would have been observed if the other treatment have been given?

We want to estimate a causal effect

The causal effect of T_i will be $y_i^1 - y_i^0$

Example: Political TV ads

What we want to say: "Individuals who watched this TV ad will be more likely to vote for Kamala Harris than if they didn't."

Estimate (estimand) of interest: ATE

- In this setup, we can't measure individual level causal effects

- **But,** we can estimate the population **average treatment effect** (ATE) by looking at those who received the treatment and those who didn't
  - $ATE = mean(y_i^1 - y_i^0)$
  - $ATE = mean(y_i^1) - mean(y_i^0)$
- Each **group** acts as a counterfactual for the other

Example: Political TV ads

What we can say: "Individuals who watched this TV ad will be more likely to vote for Kamala Harris, on average, than those who didn't."

Fundamental problem of causal inference

The fundamental problem of causal inference is that, at most, one of y_i^0 and y_i^1 can be observed at a time

- We can think of each of these as potential outcomes
  - However, we only observe one, the other is the *counterfactual*
- Estimation of causal effects requires some combination of:
  - Certain research designs that approximate potential outcomes
  - Randomization
  - Statistical adjustment

Confounders and causality

PROBLEM:

- This only works if the two groups are, on average, otherwise identical
- If the two groups differ on other factors that also cause y_i^1 and y_i^0, this is a confounding relationship
- If this is the case, our counterfactual is **wrong** and we can make no causal claim

  *If you aren't doing something to handle **all** other relevant variables (through randomization or statistical methods), you cannot make a valid causal claim, so be careful*

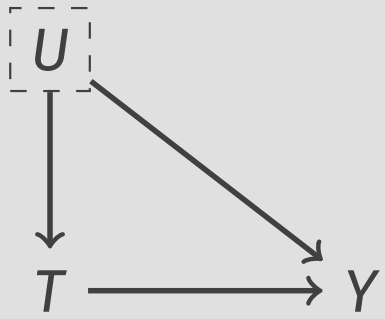
Thinking about confounding variables

(1) Direct causal relationships:

$$T \rightarrow Y$$

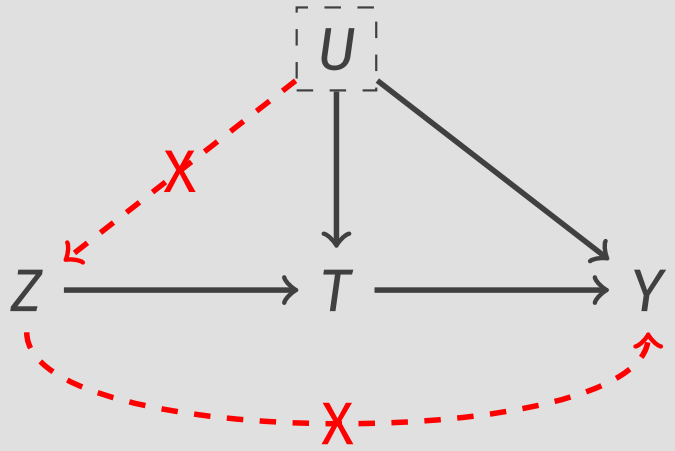
(2) Spurious relationships:

Interested in $T \rightarrow Y$, but $U \rightarrow T$ AND $U \rightarrow Y$
so T isn't related to Y, it's all U

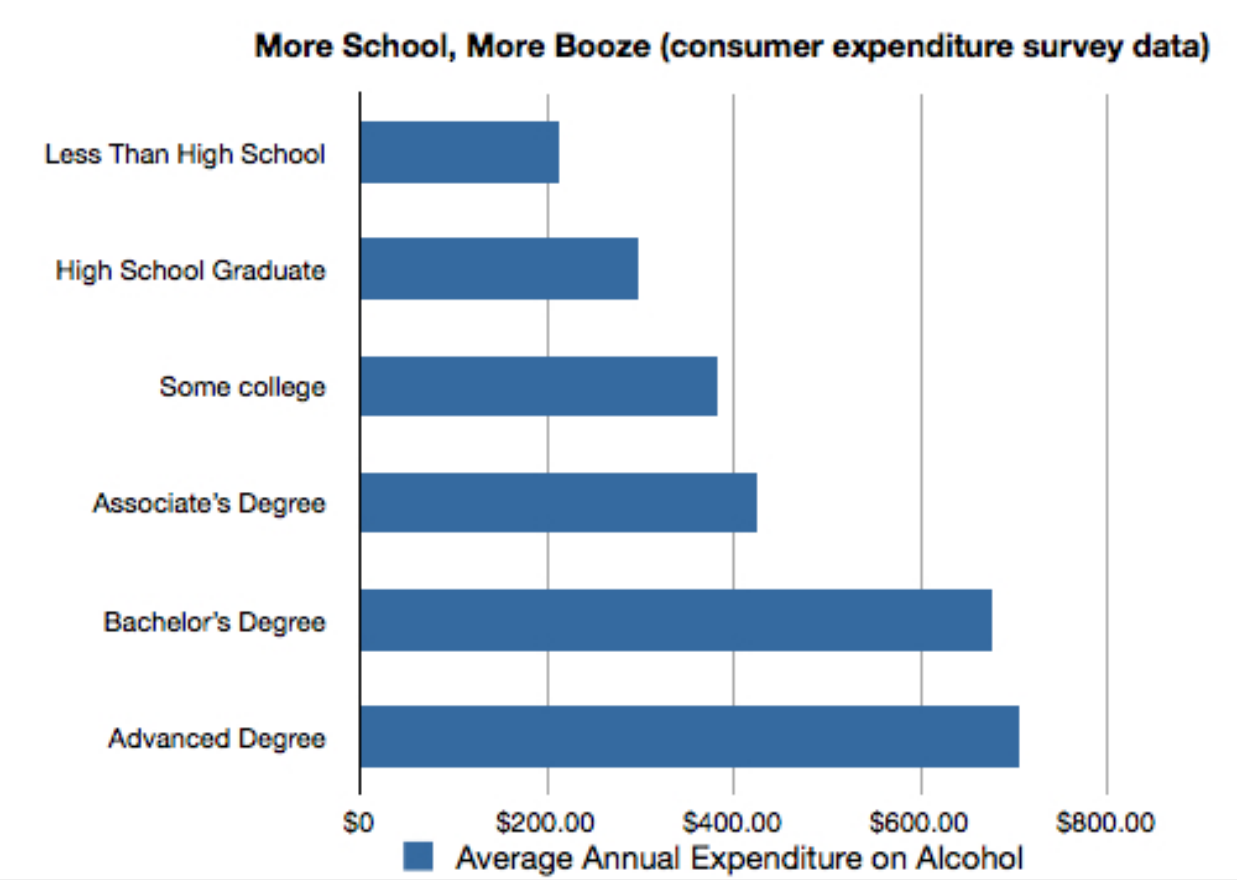


Thinking about confounding variables

(3) Chain relationships (you may hear about instrumental variables):



What's an example of each claim for this data?



- Direct causal relationships
- Spurious relationships
- Chain relationships

Being a responsible causal analyst

At a minimum we need to show ...

Association:

- What we will be doing this rest of the semester
- Correlation, contingency tables, regression coefficients, ...
- Association $\neq$ causation

Temporal order:

- For T_i to cause Y_i it must come before Y in time order
- **Post hoc ergo propter hoc**
  - "After this, therefore because of this""
- Temporal order does $\neq$ causation

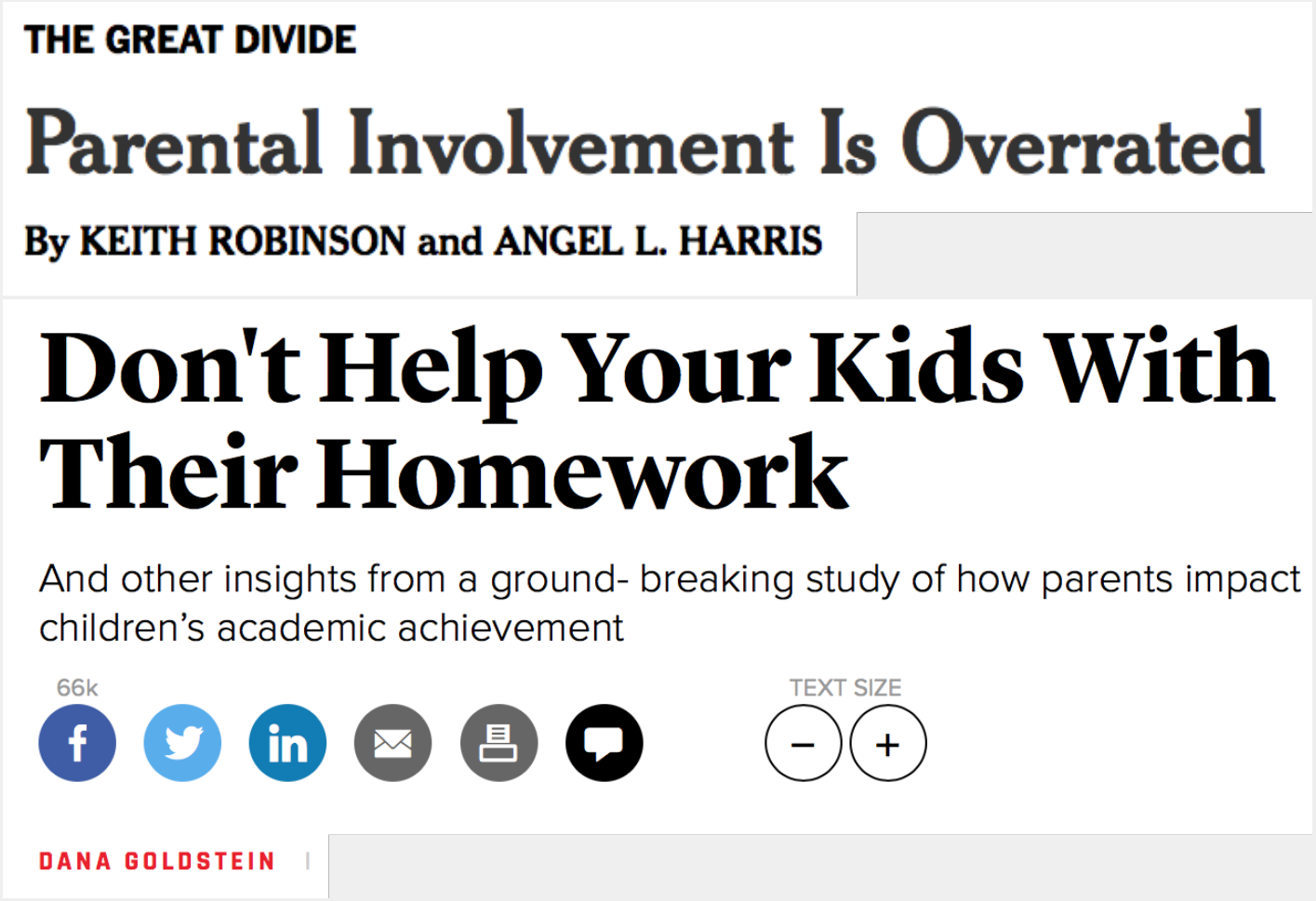
Strategy: Eliminate alternative explanations

- Suppose there is an association and a proper time order, we still cannot infer causation
- Rather, we must test for all alternative explanations
- Only if all of these have been resolved can we think about claiming causation

How can we do this?

- Experimental control
- Statistical control (Stay tuned ...)
- Research design

Examples



THE GREAT DIVIDE

Parental Involvement Is Overrated

By KEITH ROBINSON and ANGEL L. HARRIS

Don't Help Your Kids With Their Homework

And other insights from a ground- breaking study of how parents impact children's academic achievement

66k

TEXT SIZE

DANA GOLDSTEIN

Rogers, Coffman, and Bergman:

> *While the authors control for certain variables, their research only implies there is a relationship between parental involvement and student performance. This caveat is important;* ***the existence of a relationship does not tell us what causes what****.*

Think of it this way: **If you were me, and two students come to you for help, and one was getting high marks and the other low marks, which of them would you help more?** The student with low marks, right?

- An outsider, noticing that you've spent the school year helping only one of your students, might infer that instruction caused that student to earn lower grades
- This of course would not be the case, and inferring causation here would be a mistake

Example: Do Politicians Racially Discriminate?

- Is racial discrimination a problem in the acquisition of government services?
- Do legislators discriminate against individual requests for constituency service on the basis of race?

Doing it "right": Experimental Design

The sample includes states legislators in 44 U.S. states with a valid e-mail address in September 2008[2]

- Race was signaled by randomizing whether the email was signed and sent from an email account with the name Jake Mueller or DeShawn Jackson
- The text of the email was also manipulated so as to signal the partisan preference of the email sender
- The cross-tabulation between *race* & *partisan preference* gives six treatments (or groups)
- The outcome variable is the response (or lack thereof) to any e-mails

[2] Butler, D. M., & Broockman, D. E. (2011). Do politicians racially discriminate against constituents? A field experiment on state legislators. American Journal of Political Science, 55(3), 463-477.

Example: Difference of means

Table 1 Overall Effect Sizes—Does Jake Receive More Replies Than DeShawn?

	No Partisanship Signal	Republican Signal	Democratic Signal	Party Differential	
DeShawn	55.3%	54.3%	57.3%	−2.9%	Combined
Jackson	N = 806	N = 810	N = 812	*(p = 0.23)*	**−0.9%**
Jake	60.5%	56.4%	55.3%	1.1%	***(p = 0.61)***
Mueller	N = 812	N = 820	N = 799	*(p = 0.31)*	
Race Differential	?	−2.1%	1.9%		
	(p = ? *)*	*(p = 0.39)*	*(p = 0.43)*		

How do we calculate the difference between the groups?
Hypothesis testing!

Example: Calculating difference of means

```r
# create some fake data
deshawn <- rnorm(n=806, mean=55.3, sd=14.4)
jake <- rnorm(n=812, mean=60.5, sd=11.9)
```

Example: Create our TS and p-value

```r
# create function for "by hand"
# difference in means test
t.test.by.hand <- function(inputVec1, inputVec2, m0=0){
  # get the means
  m1 <- mean(inputVec1);   m2 <- mean(inputVec2)
  # standard devisations
  s1 <- sd(inputVec1);  s2 <- sd(inputVec2)
  # number of observations
  n1 <- length(inputVec1);  n2 <- length(inputVec2)
  # calculate standard error
  se <- sqrt( (s1^2/n1) + (s2^2/n2) )
  # calculate df using welch-satterthwaite df
  df <- welchDF(s1, s2, n1, n2)
  # since we have such a large sample, we'll get the same answer
  # if we use pnorm instead
  t <- (m1-m2-m0)/se
  # create dataframe with all relavent info
  dat <- c(m1-m2, se, t, 2*pt(-abs(t), df))
  # make the table pretty
  names(dat) <- c("Diff in means", "Std Error", "t", "p-value")
  return(round(dat, 3))
}
```

Example: Interpret results

```r
# you'll find this output agrees with that of t.test when
    you input x1,x2
t.test.by.hand(deshawn, jake)
```

```
Diff in means      Std Error     t            p-value
-5.165          0.647     -7.982       0.000
```

```r
# check to make sure our function works just as good
t.test(deshawn, jake, var.equal = F)
```

```
t = -7.982, df = 1524.7, p-value = 2.809e-15
alternative hypothesis: true difference in means
is not equal to 0
```

Wrap-up

Today we...

- Hypothesis testing
- Wanted to understand if X causes Y

  We talked about 2 ways to think about this:

  1. Compare two independent samples

Class business

- Read the required and suggested online materials
- Start working on problem set # 1!
- These slides are available on the course website