Political Science 15
Introduction to Research in Political Science
Lecture 5a: Intuition behind Hypothesis Testing

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Course review - what we've learned so far

- We want to know about relationships in the world: does smoking cause cancer? do vitamins increase lifespan? does democracy cause development? does HCL cure COVID-19?
- First, we talked about how, even though we see a correlation between two variables, it's not necessarily a causal relationship.
- Next, we talked about how exogeneity, usually through random assignment, allows us to make a causal claim.
- Then, we discussed the building blocks for probability and uncertainty. This is where Lecture 3 pays off!
- In Lecture 4, we learned how we can formally detect relationships through linear regression, using a model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ which we estimate.

Today

Now, we move on to ask **whether the relationship we observed is real or just due to chance**.

Say we estimate a certain $\hat{\beta}_1$ which implies that an additional year of schooling is associated with \$1,000 more in annual wage.

So, the slope of our regression line is positive. Great, right? Not so fast! This could be due entirely to chance.

In this lecture, we will learn to detect **statistically significant** effects.

Important

Just because an effect is statistically significant, does not mean it is **causal**. All the same rules from Lecture 2 on deciding whether a causal claim is warranted still apply!

Motivation for hypothesis testing

Scientists want to **test hypotheses** about patterns they observe in the data.

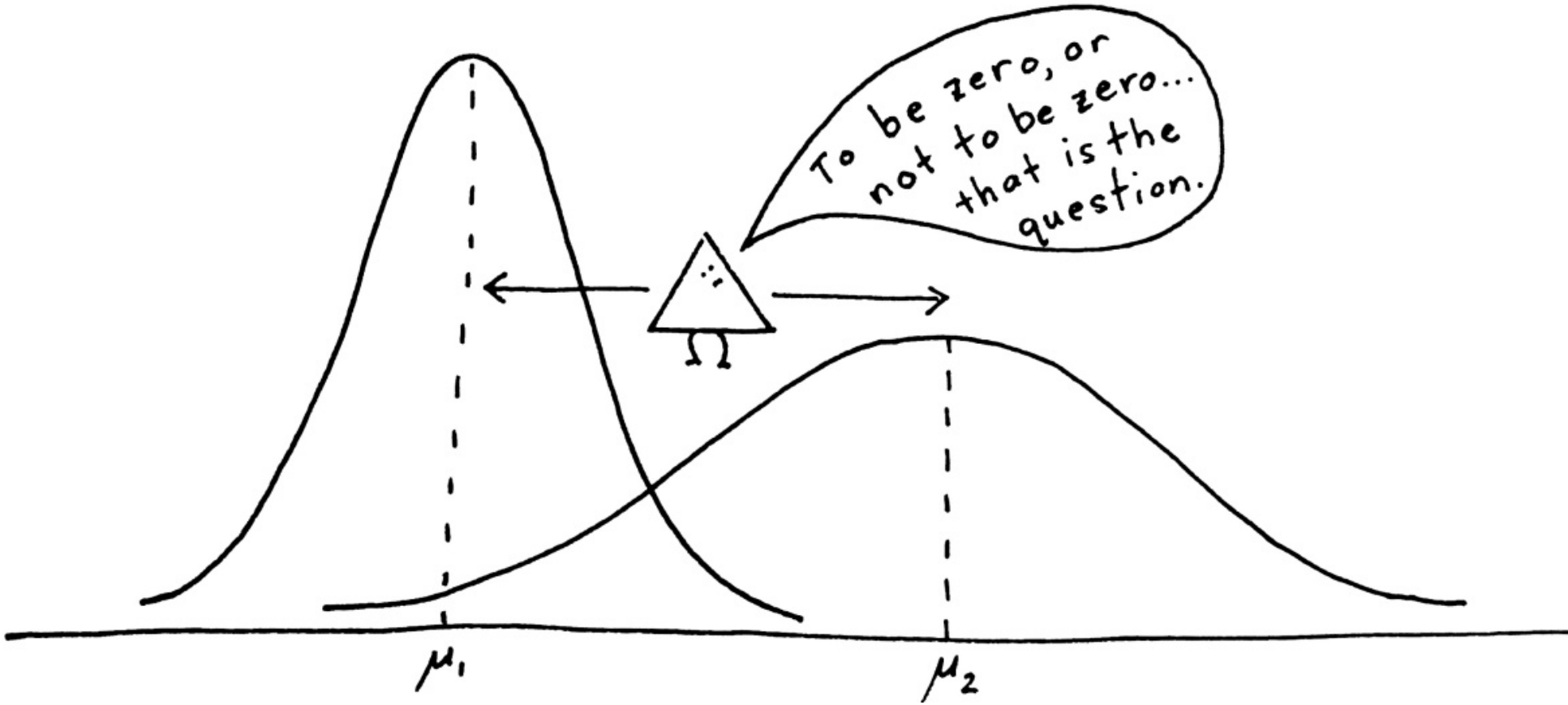
Remember the multiverse? If you think probabilistically, you *could have* observed other (more or less likely) outcomes in the data.

How can you be sure the pattern you observe is real? We are data detectives.

We want to investigate whether the pattern we see if a signal, or if it is just noise.

A Shakespearean question

The philosophical question asked by the little Δ will make sense at the end of the lecture!
Reminder: read **Chapter 4** of Real Stats!



Artwork by allison_horst

First, let's talk about beer

- In 1908, William Sealy Gosset was the "Head Experimental Brewer" at the Guinness Brewing Company in Dublin, Ireland.
- He wanted to experiment on different barley varieties to see how they impacted the quality of stout (i.e. beer).
- Sealy Gosset came up with the statistical test called Student's t-test.
- Specifically, he came up with the t-distribution which help us make inferences in small samples (more on that later).

Two-sample tests

Definition

Two-sample tests allow us to test whether the difference between two *populations* is statistically significant.

We can compare one *sample* to another *sample* and answer the question: "Are these samples from the same population or from different populations?".

Key intuition behind hypothesis testing: We want to understand how likely we are to observe some difference if there really is no difference.

This requires us to think about how "weird" our result is, if there really is no difference at all!

Null hypothesis

First, we need to define our null hypothesis, which is typically a hypothesis of no effect (e.g. the beer samples are not different, the effect of schooling on wage is 0, etc.).

Check. If two samples are drawn from populations that have the same mean, then in theory, what do you expect the difference between the two sample means to be? The difference in means would be 0, if the samples came from populations with the same mean.

This is how we would state our **null hypothesis** formally.
H_0: the means are equal (the difference in means $= 0$)

Now we need to specify an **alternative hypothesis**.
H_A: the means are not equal (the difference in means $\neq 0$)

The intuition behind hypothesis testing

How do we adjudicate between our two competing hypotheses?

We think probabilistically (recall Lecture 3) about what the state of the world would be if the null hypothesis were true.

Use your best Hollywood trailer voice from the 90s:
"In a world" ...dum, dum, dum... "where the null hypothesis is true".

This "null state of the world" will have a range of possible outcomes which are more or less likely. This is the **null distribution**.

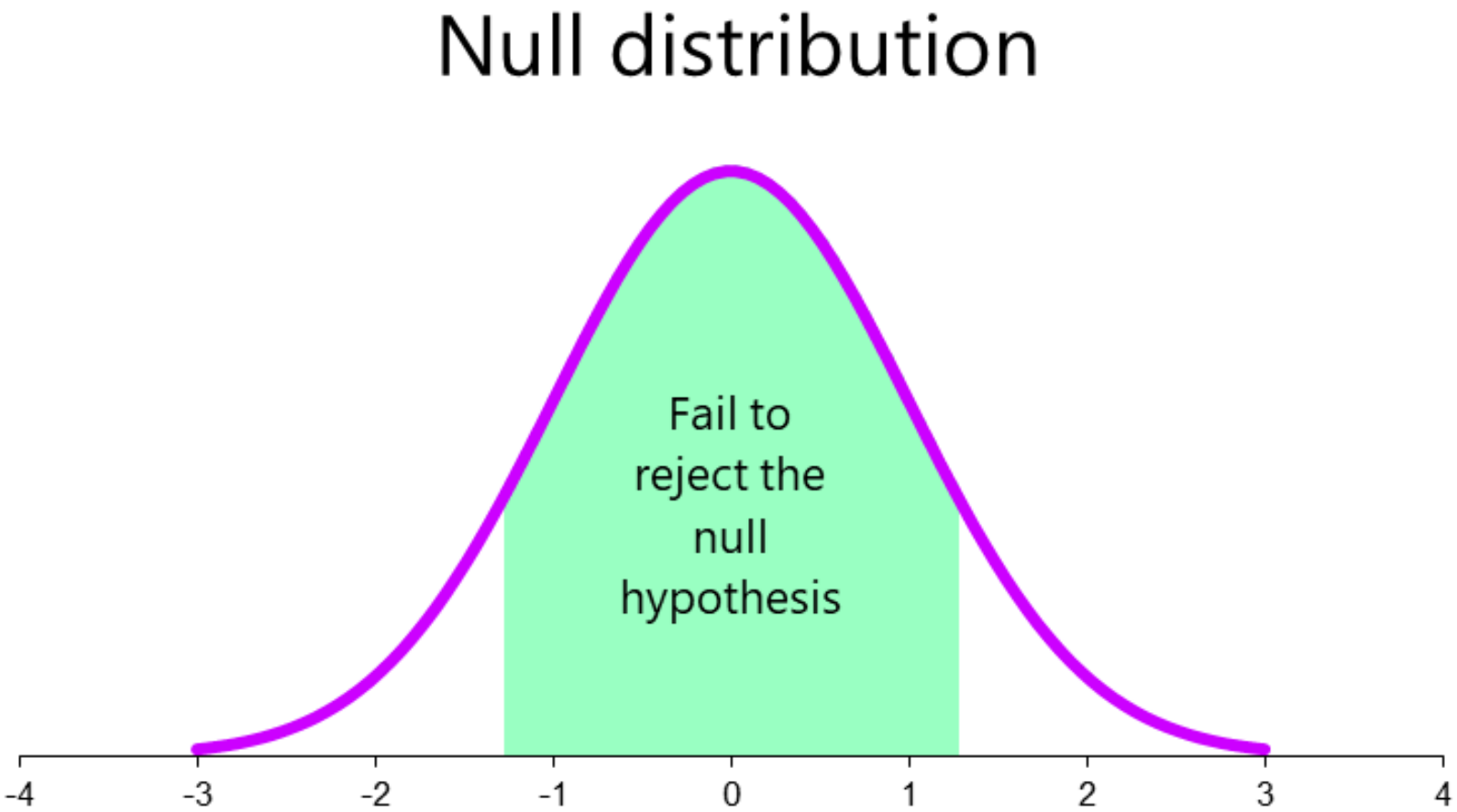
We then think about *how likely* we would be to observe the outcome that we did, *if* the null were true.

If it is highly unlikely, we can reject the null hypothesis as an unreasonable state of the world.

Thus, we can conclude that the outcome/effect we observed is unlikely to be due to chance alone, and is hence statistically significant.

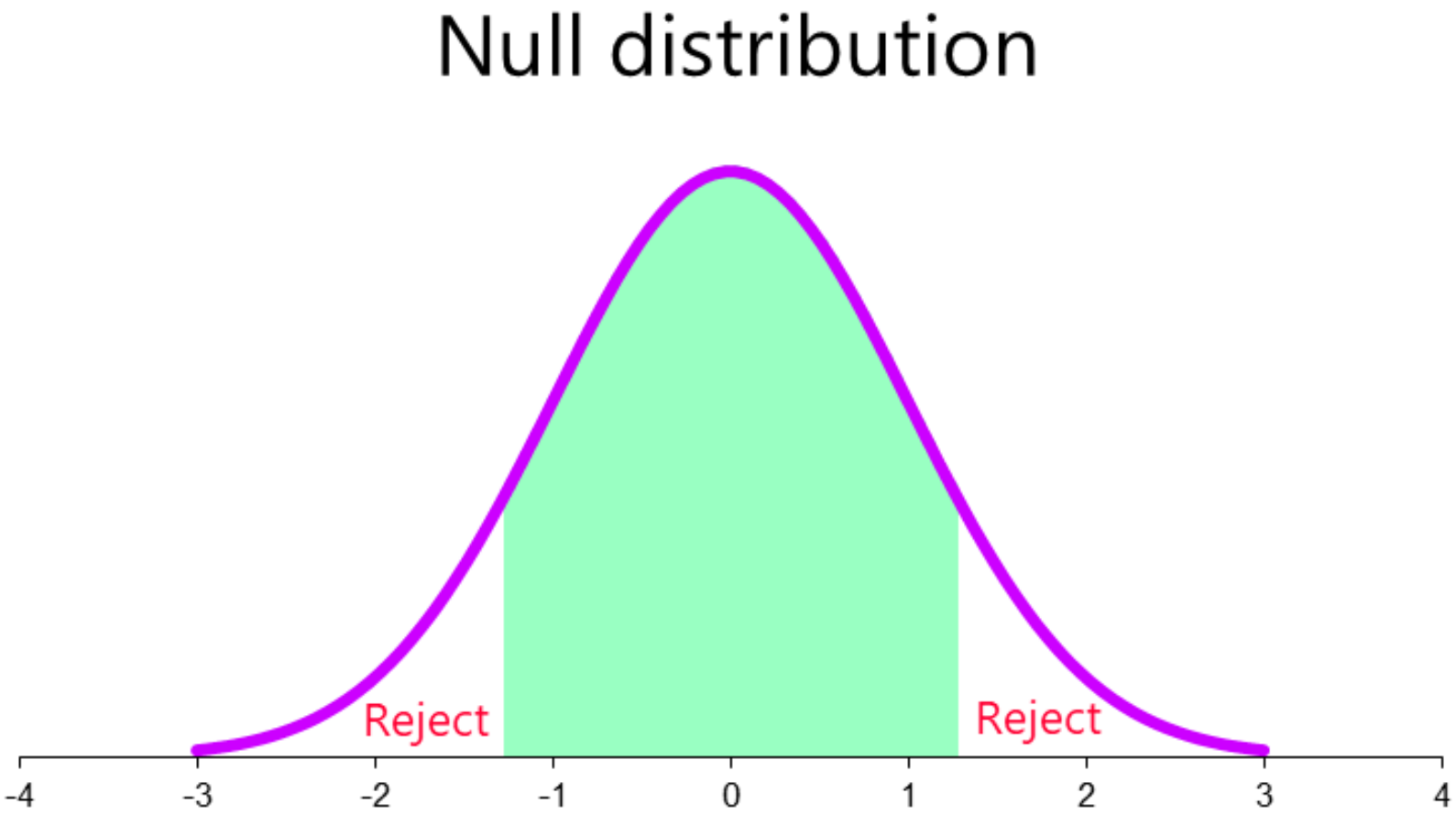
The intuition behind hypothesis testing

If we are in the green, the outcome we observed would have been pretty *likely* if the null hypothesis were true, so we **fail to reject the null hypothesis**.



The intuition behind hypothesis testing

If we are in the white, the outcome we observed would have been pretty *unlikely* if the null hypothesis were true, so we **reject the null hypothesis** (in favor of the alternative).



The intuition behind hypothesis testing

This is the intuition behind hypothesis testing: we ask how "weird" our observed outcome is in a world where the null is true, and on that basis decide if the null is an unreasonable state of the word (i.e. we reject H_0), or if it is not that weird (i.e. we fail to reject H_0).

We need to add more bells and whistles to this statistical machinery, namely:

- How do we measure the outcome, i.e. what statistic do we use? (E.g. Z-score, t-statistic)
- What distribution do we use for the null? (E.g. normal, t)
- Do we need 1 or 2 rejection regions? (I.e. one or two-sided H_A)
- How do we decide where to place the fences? (I.e. What critical values do we use? What significance level?)

And... we will finally learn what the mysterious "t-statistics" and "p-values" in our regression table mean!

Watch out for this common mistake

If the outcome we observe is pretty likely if the null hypothesis were true, we fail to reject the null hypothesis.

We **never "accept" the null hypothesis**! This just means that we haven't found enough evidence to reject the null of *no effect*.

Think of a "not guilty" verdict rendered by a jury. The accused *may* be guilty (i.e. the jury may be making a *mistake*), but the evidence was not sufficient to convict.

Political Science 15
Introduction to Research in Political Science
Lecture 5b: How to conduct hypothesis testing: two-sample tests

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Two-sample tests

Let us now go back to two-sample tests.

The quantity of interest here is the **difference in means** between two groups on some outcome, for example:

- difference in "tastiness" of Guinness beer depending on which type of barley it is brewed with
- difference in voter turnout among two groups of people (e.g. Republicans vs. Democrats)
- difference in probability of war in two groups of countries (e.g. autocracies vs. democracies)

Two-sample test: political science example

Our running example will be *vote intention*.

- We survey 60 people from campus, asking whether they intend to vote in the next election.
- 25 reported being Republican, 35 reported being Democrat, which gives us two groups.
- We'll get a mean vote intention for both, which we'll call $\overline{Vote}_R$ and $\overline{Vote}_D$.
- Check: why is there a line over these quantities?
  - They are sample means, just like $\bar{X}$.
- We expect $\overline{Vote}_R$ won't exactly equal $\overline{Vote}_D$, but we want to know if the difference is just due to *chance*, or if it is a *real* difference.

Let us test this

1. First, we need to **state our null and alternative hypotheses**.

What is our null hypothesis here?

- *in words*: there is no difference in the average vote intention between Republicans and Democrats
- *in math*: $H_0 : \overline{Vote}_R = \overline{Vote}_D$

What is the alternative hypothesis (two-sided)?

- *in words*: there is some difference in the average vote intention between Republicans and Democrats
- *in math*: $H_A : \overline{Vote}_R \neq \overline{Vote}_D$

Let us test this

1. Second, we need to **estimate a statistic**, in this case, the difference in means (DIM).

We find that:

- among Republicans, 19/25 report that they will vote ($\overline{Vote}_R = 0.76$)
- among Democrats, 22/35 ($\overline{Vote}_D = 0.63$) say they will vote

$\Rightarrow$ $DIM = \overline{Vote}_R - \overline{Vote}_D = 0.13$

1. Third, we need to **derive the null distribution**, i.e. the distribution of possible outcomes if the null hypothesis were true.

We will do this formally later on. But first, a thought experiment. Consider the tale of the sloppy research assistant.

The tale of the Sloppy Research Assistant

Imagine an Excel sheet for this dataset:

- one row per person
- in the first column, we record *Vote* equals 0 or 1 to indicate whether that person says they will vote
- in the second column, we record *R* or *D*

ID	Vote	Party
1	1	R
2	0	D
⋮	⋮	⋮
60	1	D

You compute a mean of *Vote* for those with *R*, a separate mean of *Vote* for those with *D*, and subtract:

$$DIM = \overline{Vote}_R - \overline{Vote}_D = 0.13$$

Coding note: how would you do this in R? See Section 2 material.

```
mean(dat$Vote[dat$Party == "R"]) - mean(dat$Vote[dat$Party == "D"])
```

The tale of the Sloppy Research Assistant

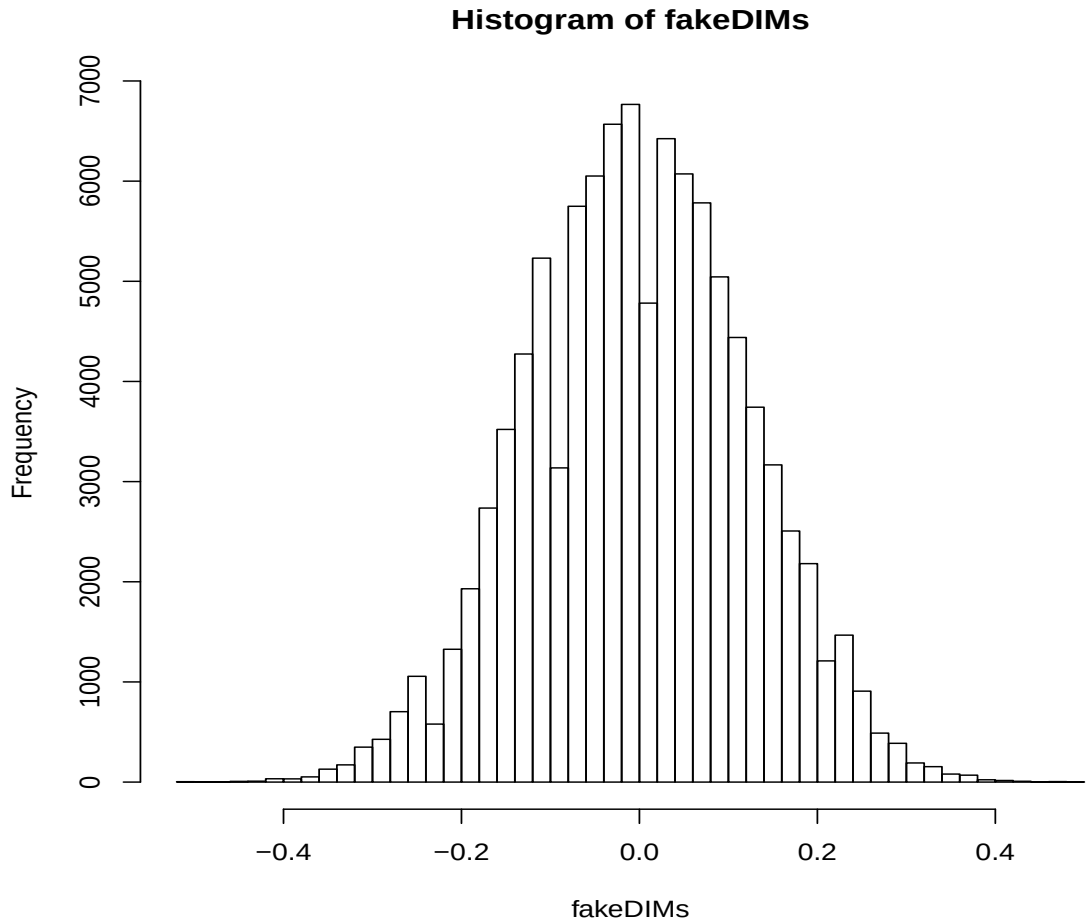
Now, suppose the RA accidentally scrambled the order of the second column:

- What do you expect to see for your difference in means using these "wrong" data?
- What if you rescrambled it, still getting the wrong order, would you see the same thing?
- So if you rescramble it and get the DIM many times, what do you obtain?

The distribution of our scrambled data

So if we look at the distribution of our fake, scrambled data what do we have? We'll call this our "Fake difference in means (DIM)".

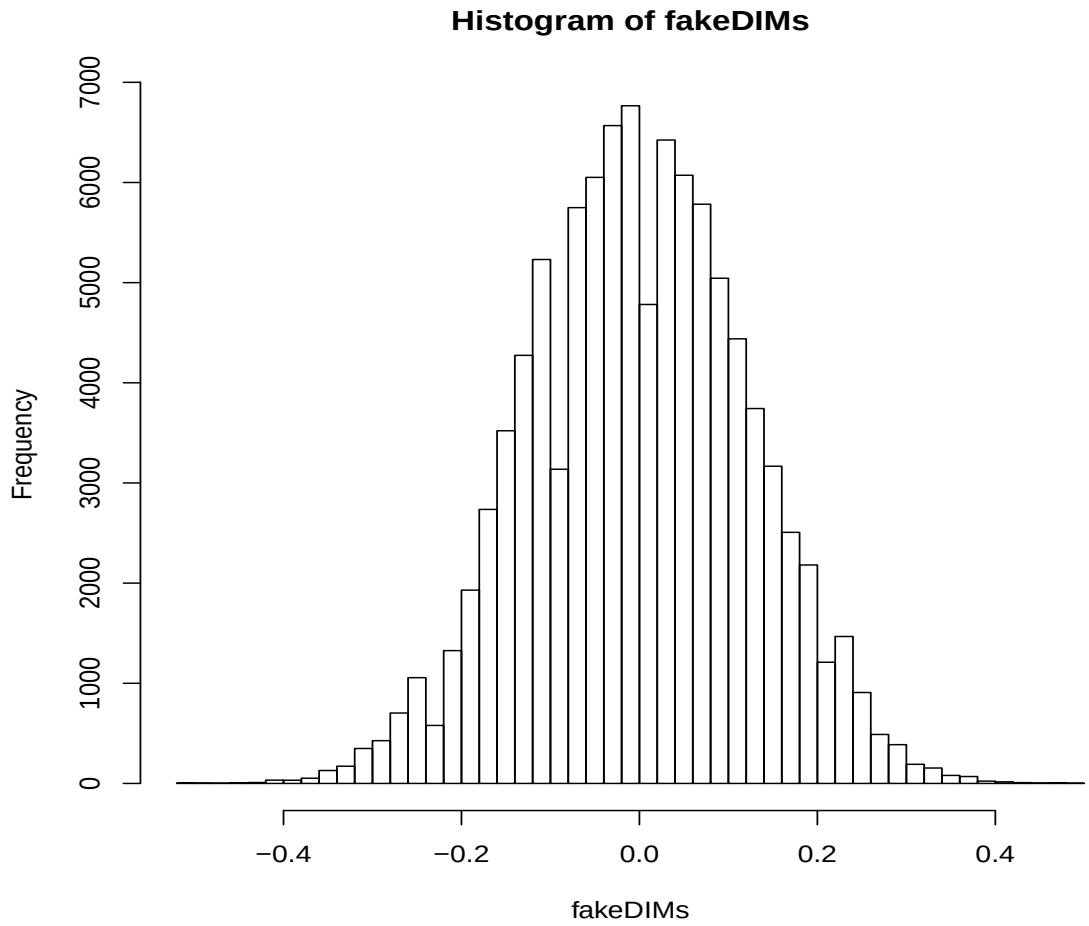
Remember, each time we are taking a sample and then computing the difference in mean between the Dems and Reps.



What to note about the distribution of our scrambled data

What does the shape of the distribution look like?

What value is it centered on?



What to note about the distribution of our scrambled data

What does the shape of the distribution look like? Like a normal distribution.

Why? Hint: one of the two awesome theorems from Lecture 3.
The Central Limit Theorem! This is so cool! Because of the CLT, we are guaranteed that this distribution will always be normal. We don't need to simulate the multiverse any more!

What value is the distribution centered on?
Centered on 0.

$\Rightarrow$ This is a simulated **null distribution** centered on the expected value if the null hypothesis were true (H_0: no difference in voting intention between 2 groups).

How "weird" is "weird"?

How do we decide that our observed DIM is super weird, if the null were true?

Just count up the proportion of outcomes more extreme than ours under the null:

```
>right_tail=sum(fakeDIMs>=DIM)
>left_tail=sum(fakeDIMs<= (-DIM))
>pval = (right_tail + left_tail)/iters
>pval
[1] 0.17865
```

This number is the probability of observing an outcome as extreme as we did if the null were true.

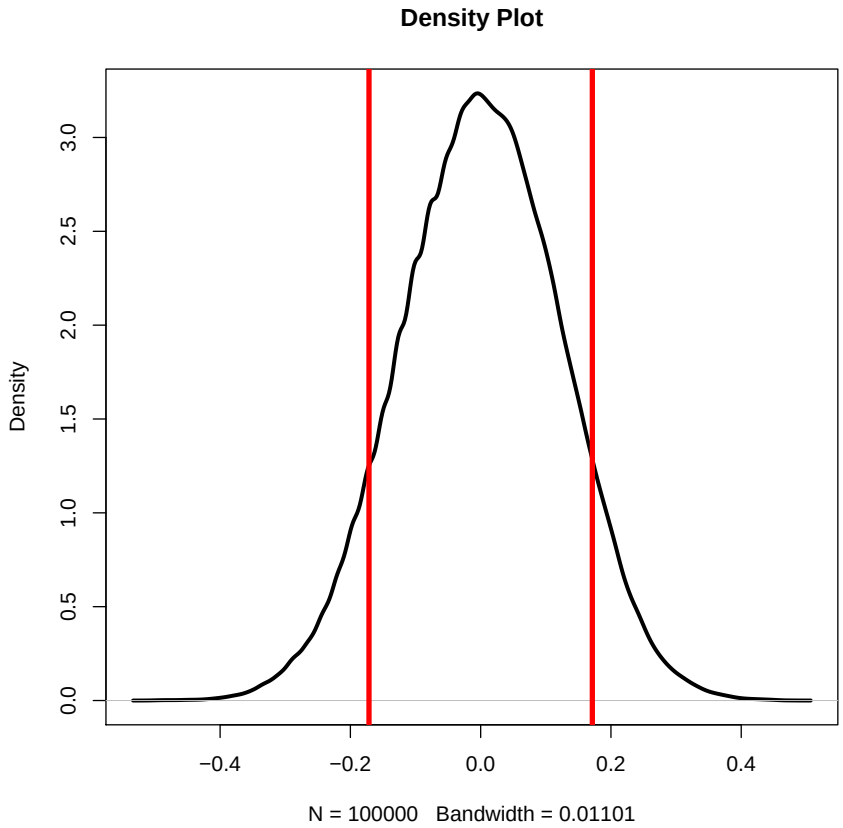
This is the definition of a p-value!

How do we interpret this p-value?
We fail to reject the null hypothesis that the DIM is 0, that is, we cannot reject the null that the vote intentions $\overline{Vote}_R$ and $\overline{Vote}_D$ are from populations with the same mean.

Graphical representation of p-values

Let's replot our histogram of scrambled data, but now using a continuous approximation:

```
plot(density(fakeDIMs), main="Density Plot")
abline(v=DIM, col=2, lwd=4)
abline(v=-DIM, col=2, lwd=4)
```



One-sided versus two-sided test

Let's add one of the bells and whistles to our analysis.

✓ Do we need 1 or 2 rejection regions? (I.e. one or two-sided H_A)

In this case, our alternative hypothesis was "two-sided":

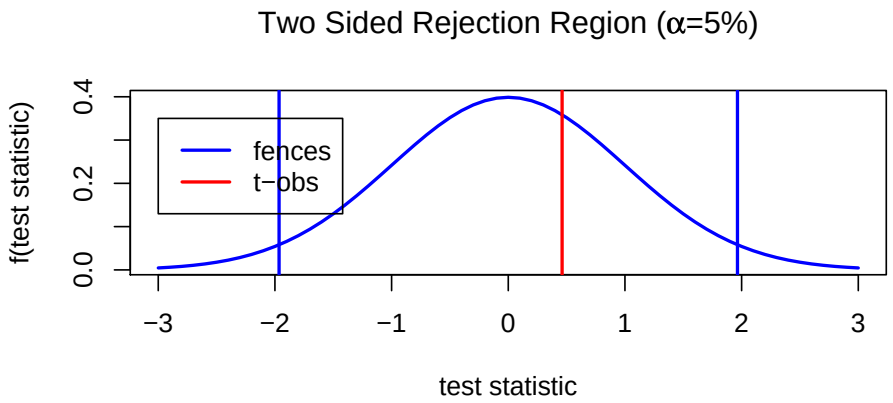
- *in words*: there is some difference in the average vote intention between Republicans and Democrats
- *in math*: $H_A : \overline{Vote}_R \neq \overline{Vote}_D$

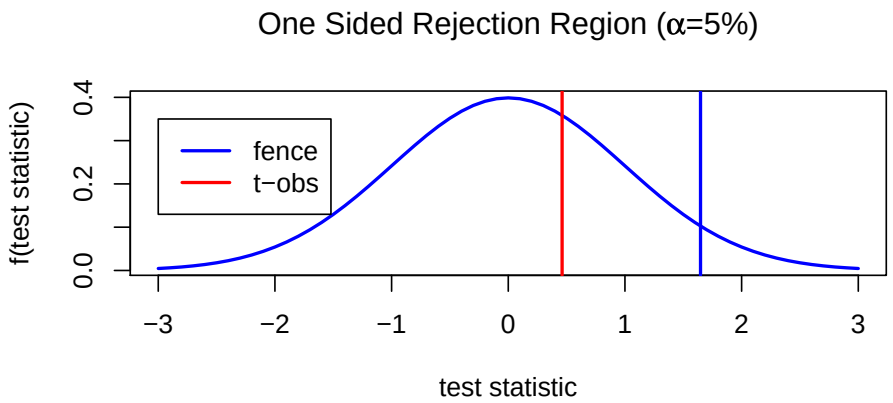
Instead, we could use a "one-sided" alternative hypothesis:

1. Option 1.
   - *in words*: Republicans have, on average, a higher vote intention than Democrats
   - *in math*: $H_A : \overline{Vote}_R > \overline{Vote}_D$ or $H_A : DIM > 0$
2. Option 2.
   - *in words*: Republicans have, on average, a lower vote intention than Democrats
   - *in math*: $H_A : \overline{Vote}_R < \overline{Vote}_D$ or $H_A : DIM < 0$

One-sided versus two-sided tests

- Two-sided: rejection region is in **both** tails
- One-sided: rejection region is in **one** of the tails





In practice, we use **two-sided tests**. One-sided tests are not appropriate unless we have strong theoretical priors on the direction of the effect.

Back to p-values

Key definition

A p-value is the probability of observing an outcome (e.g. a DIM or a coefficient) as extreme as we did if the null hypothesis were true.

To conclude the tale of the sloppy RA, we decided that a p-value of ≈ 0.18 was still a pretty likely outcome ($\approx$ 4 in 20 cases) if the null were true.

But how do we decide how unlikely the observed outcome has to be before we can reject the null?

1. First, decide on a one- or two-sided test.
   - Let's go with a two-sided test.
2. Second, decide on a **significance level**, which states how unlikely an outcome has to be for us to reject the null H_0.
   - Let's go with $\alpha = 0.05$, or 1 in 20 cases.

Significance level α

Bell and whistle #2

- ✓ How do we decide where to place the fences? (I.e. What critical values do we use? What significance level?)

When we pick a significance level of $\alpha = 0.05$:

- We have a 5% probability of observing the effect due to chance alone.
- We're okay with that. $\alpha = 0.05$ is the conventional level used for determining statistical significance.

Easy steps to remember

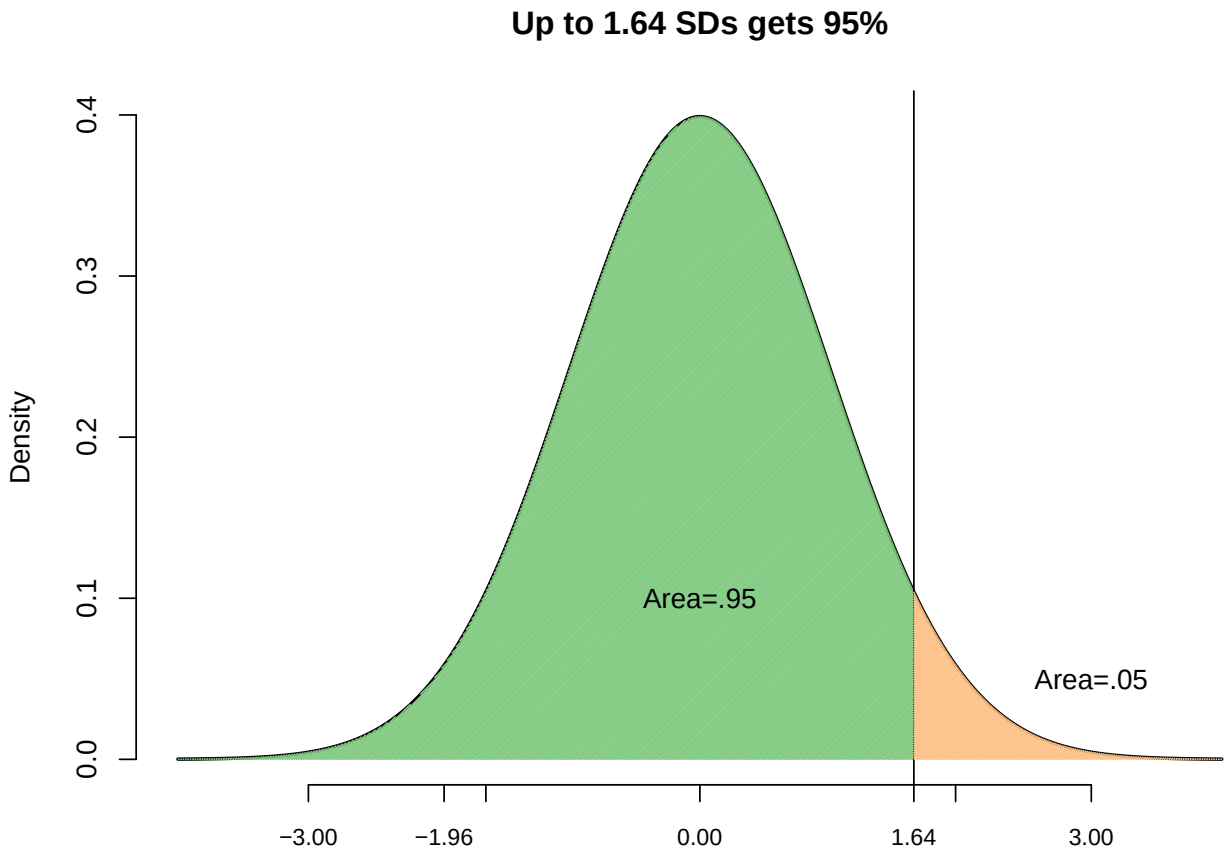
- If the p-value is less than 0.05, you reject the null hypothesis.
- If the p-value is greater than 0.05, you fail to reject the null hypothesis.

Here, we compare our computed p-value to the threshold of our chosen significance level α.

Standard normal distribution

When we have $\mathcal{N}(0, 1)$, we call it the "standard normal" distribution.
Check: what does this mean? This is a normal distribution with a mean of 0 and a standard deviation of 1.

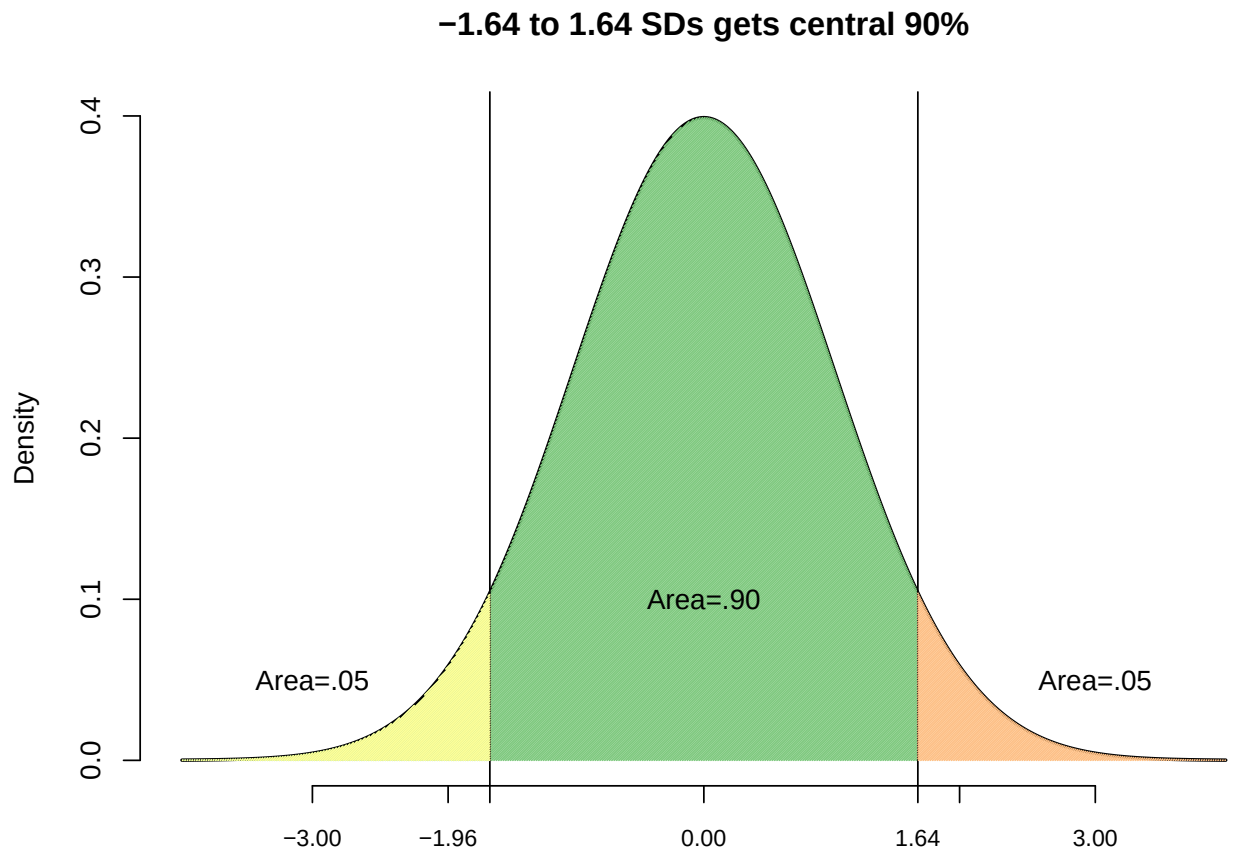
Two key **critical values** from this distribution are 1.64 and 1.96.



Standard normal distribution

When we have $\mathcal{N}(0, 1)$, we call it the "standard normal" distribution.
Check: what does this mean? This is a normal distribution with a mean of 0 and a standard deviation of 1.

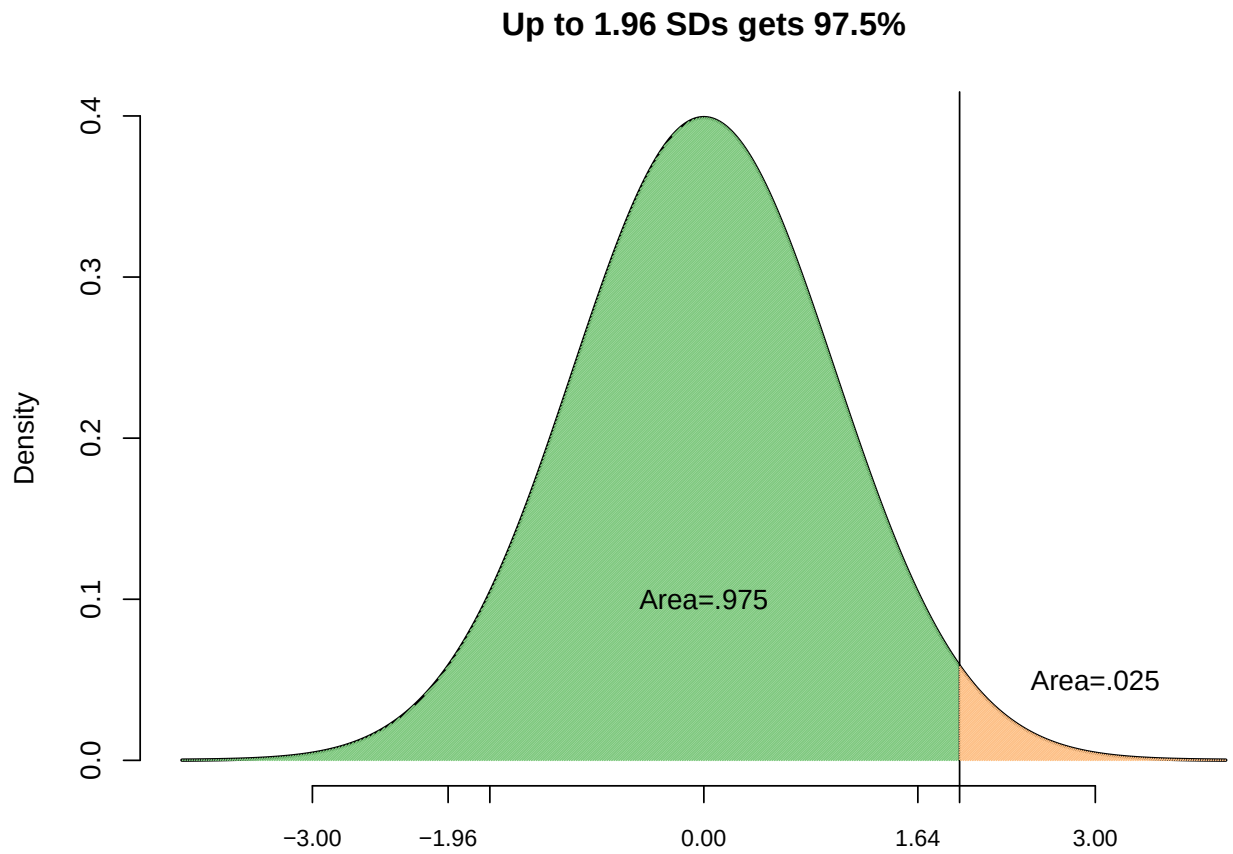
Two key **critical values** from this distribution are 1.64 and 1.96.



Standard normal distribution

When we have $\mathcal{N}(0, 1)$, we call it the "standard normal" distribution.
Check: what does this mean? This is a normal distribution with a mean of 0 and a standard deviation of 1.

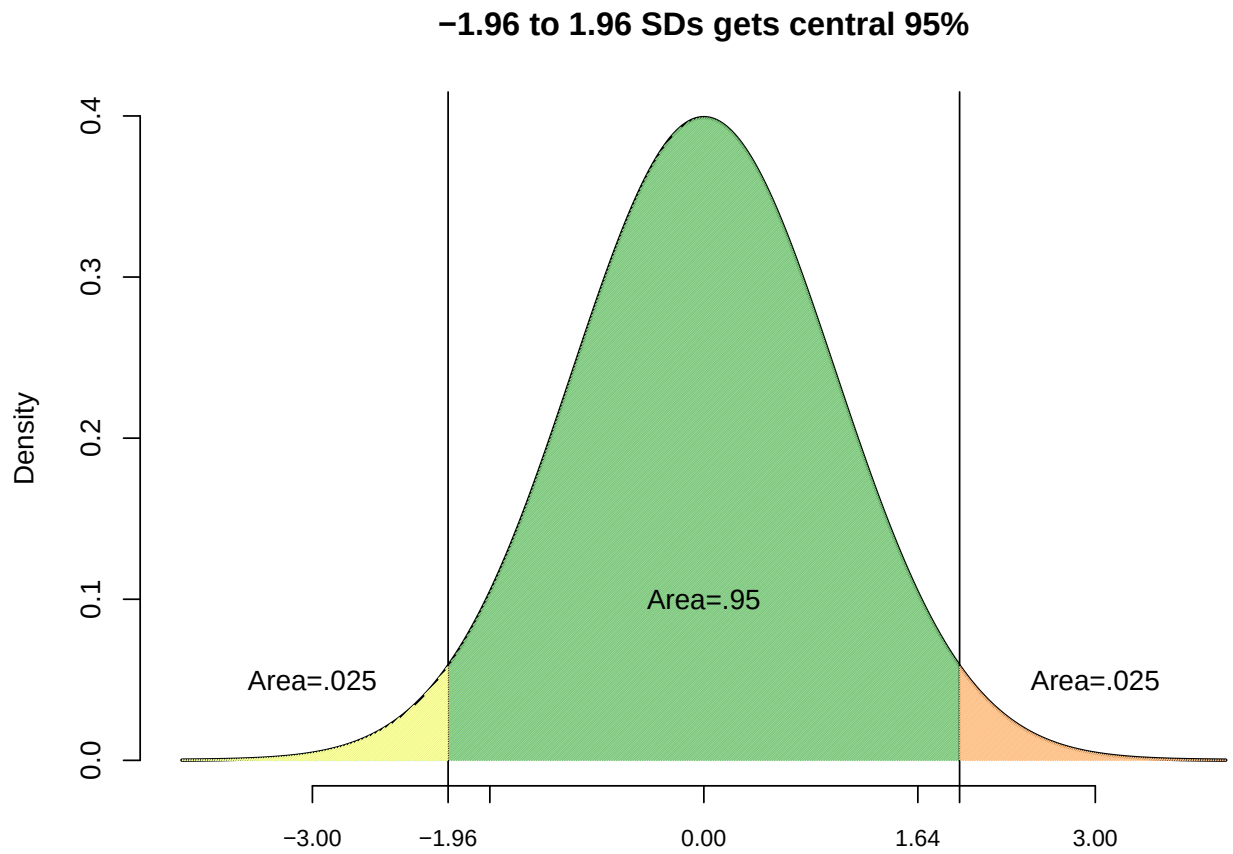
Two key **critical values** from this distribution are 1.64 and 1.96.



Standard normal distribution

When we have $\mathcal{N}(0, 1)$, we call it the "standard normal" distribution.
Check: what does this mean? This is a normal distribution with a mean of 0 and a standard deviation of 1.

Two key **critical values** from this distribution are 1.64 and 1.96.



Another way to place the fences

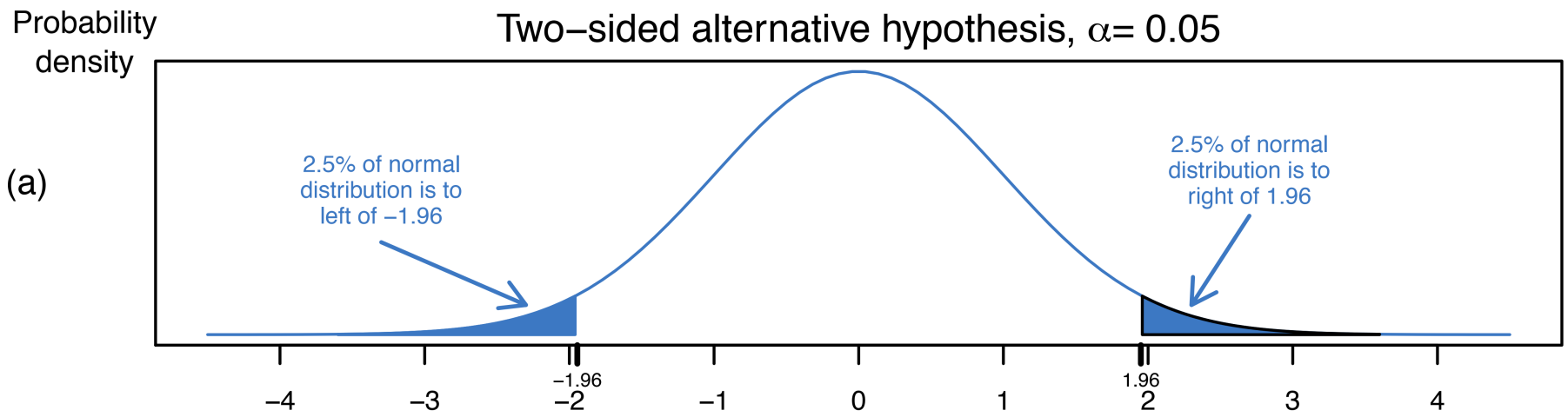
We could compare our DIM to the threshold of our chosen critical value c.

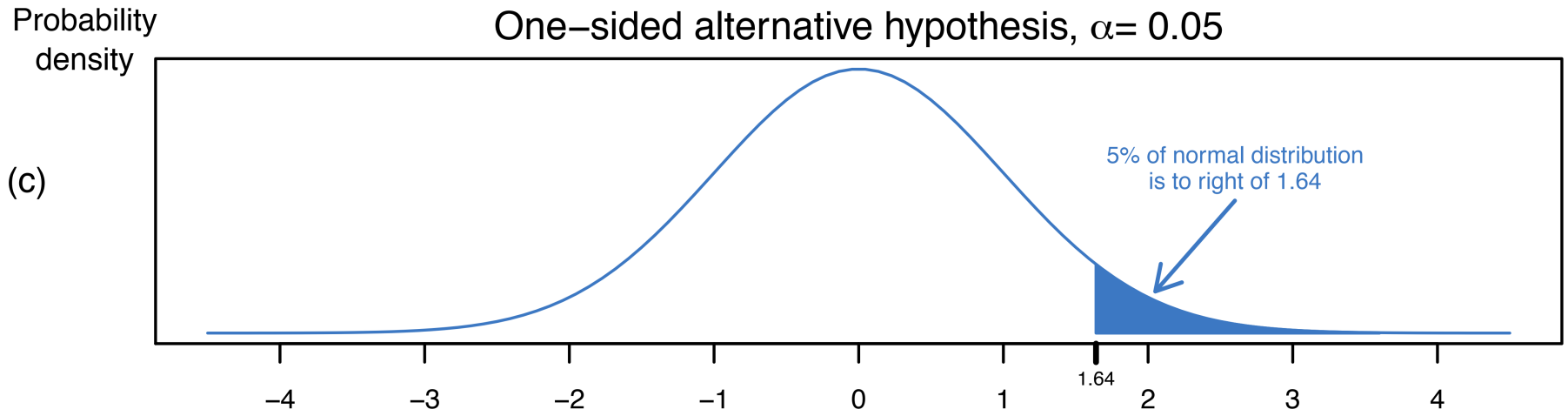
The critical value is how many deviations away from 0 our observed outcome is. (If it is many deviations away, it is pretty unlikely).

1. First, choose between a one-sided or two-sided test. Again, let's go for two-sided.
2. Second, choose a critical value c above which the observed effect would be so unlikely that we reject the null.
   - For example, let's choose $c = 1.96$.
   - This means our effect would need to be 1.96 standard deviations away from 0 (i.e. super unlikely) for us to reject the null hypothesis of no effect.
   - The critical value is the threshold. What do we compare against it? A test statistic (more on that later).

Relationship between critical value and significance level

There is a relationship between the critical value c and the significance level α. Both can be used as thresholds.



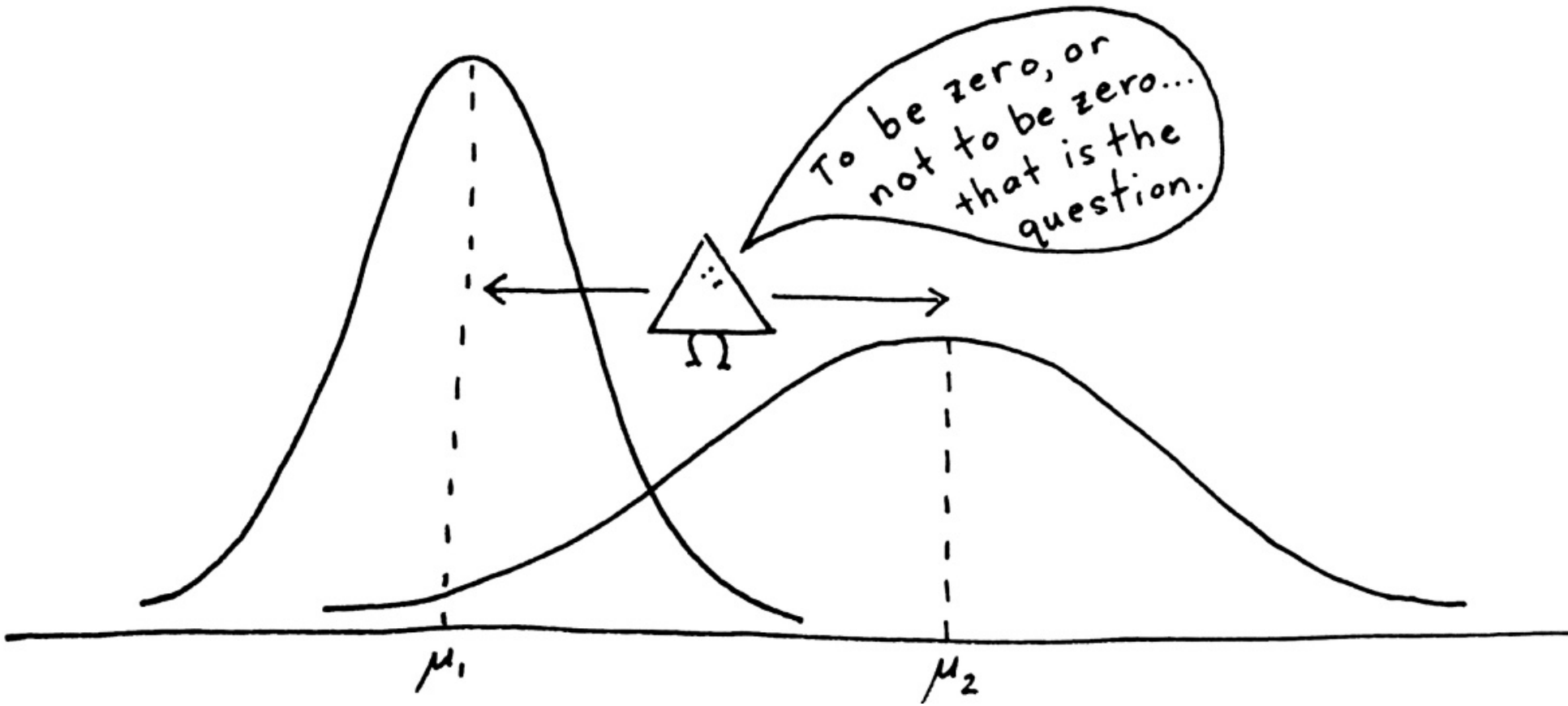


Review so far

- We want to test whether an observed effect is statistically significant, or if it is due to chance alone.
- For example, whether Republicans and Democrats differ significantly in their intention to vote.
- We formally state our null and alternative hypotheses.
  - Usually, the null hypothesis H_0 is of *no* effect (i.e. no difference).
  - Usually, the alternative hypothesis H_A is two-sided, i.e. there is *some* non-zero effect (i.e. some difference).
- Next, we need to determine just how "weird" our observed effect is, "in a world" where the null hypothesis is true. That is, how unlikely the observed outcome is in the null distribution.
  - If it is not that weird, we fail to reject the null. It is a plausible outcome "in a world" of no effect.
  - If it is very weird, we reject the null hypothesis, as there is only a tiny chance that the effect occurred due to chance alone.
- We use p-values and critical values as thresholds to determine what is likely and what is unlikely.

Remember this picture?

Applied to our vote intention example, the population mean for Democrats is μ_1 and the population mean for Republicans is μ_2. We use data to estimate the sample equivalent $\overline{Vote}_D$ and $\overline{Vote}_R$. Δ is the true difference, which we estimate with DIM. Finally, we test $H_0 : \Delta = 0$.



Next time

We just learned the basic statistical machinery to do hypothesis testing.

Next, we add the remaining bells and whistles:

- How do we measure the outcome, i.e. what test statistic do we use? (E.g. Z-score, t-statistic)
- What distribution do we use for the null hypothesis? (E.g. normal, t)

We will also cover:

- How to conduct hypothesis testing in the context of regression
- How to do this in R
- Type I and Type II errors
- Confidence intervals

Political Science 15
Introduction to Research in Political Science
Lecture 5c: Hypothesis Testing with Regression Analysis

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Review of previous modules

- We want to **test whether an observed effect is statistically significant**, or if it is due to chance alone.
- For example, whether Republicans and Democrats differ significantly in their intention to vote.
- We **formally state our null and alternative hypotheses**.
  - Usually, the null hypothesis H_0 is of *no* effect (i.e. no difference).
  - Usually, the alternative hypothesis H_A is two-sided, i.e. there is *some* non-zero effect (i.e. some difference).
- Next, we need to determine just how "weird" our observed effect is, "in a world" where the null hypothesis is true. That is, **how unlikely the observed outcome is in the null distribution**.
  - If it is not that weird, we fail to reject the null. It is a plausible outcome "in a world" of no effect.
  - If it is very weird, we reject the null hypothesis, as there is only a tiny chance that the effect occurred due to chance alone.
- We use **p-values** and **critical values** as thresholds to determine what is likely and what is unlikely.

Review of previous modules

We developed our intuition for the machinery of hypothesis testing. We also talked about the following bells and whistles:

- ✓ Do we need 1 or 2 rejection regions? (I.e. one or two-sided H_A)
- ✓ How do we decide where to place the fences? (I.e. What critical values do we use? What significance level?)
- ☐ How do we measure the outcome, i.e. what test statistic do we use? (E.g. t-statistic)
- ☐ What distribution do we use for the null? (E.g. normal, t)

Today

Today, we'll discuss

- How to conduct hypothesis testing in the context of regression
- How to do this in R
- Type I and Type II errors

We'll cover the other two "bells and whistles":

- What **test statistic** to use to measure the effect.
- What **distribution** to use for the null hypothesis.

Hypothesis testing with regression

Recall that in our previous modules, we conducted hypothesis testing using **two-sample tests**.

- That is, we asked whether two samples were statistically significantly different from each other.
- The quantity of interest was the Difference in Means (DIM).

Today, we will conduct hypothesis testing in the context of **regression analysis**.

- Here, the quantity of interest will be slope coefficient β_1.

Hypothesis testing with regression

- We estimate a bivariate regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.
- We want to find out if the relationship between X and Y is **real**, or if it is just **by chance**?
- We want to answer the question "Is the *association* between X and Y statistically significant, or did we just get lucky?"
- Remember the multiverse. It could be that the *sample* we drew showed a relationship between X and Y. But with a slightly different sample, maybe we would not have observed any relationship after all.

Stating our hypotheses in regression

1. First, we need to **state our null and alternative hypotheses**.

What is the null hypothesis in this context?

- *in words*: there is no association between X and Y (or X has no effect on Y)
- *in math*: $H_0 : \beta_1 = 0$
- Check. Why is there no hat on β_1? Because we want to make inferences on the underlying *population* parameter!

What is the alternative hypothesis (two-sided)?

- *in words*: there is some association between X and Y (or X has some effect on Y, whether positive or negative)
- *in math*: $H_A : \beta_1 \neq 0$

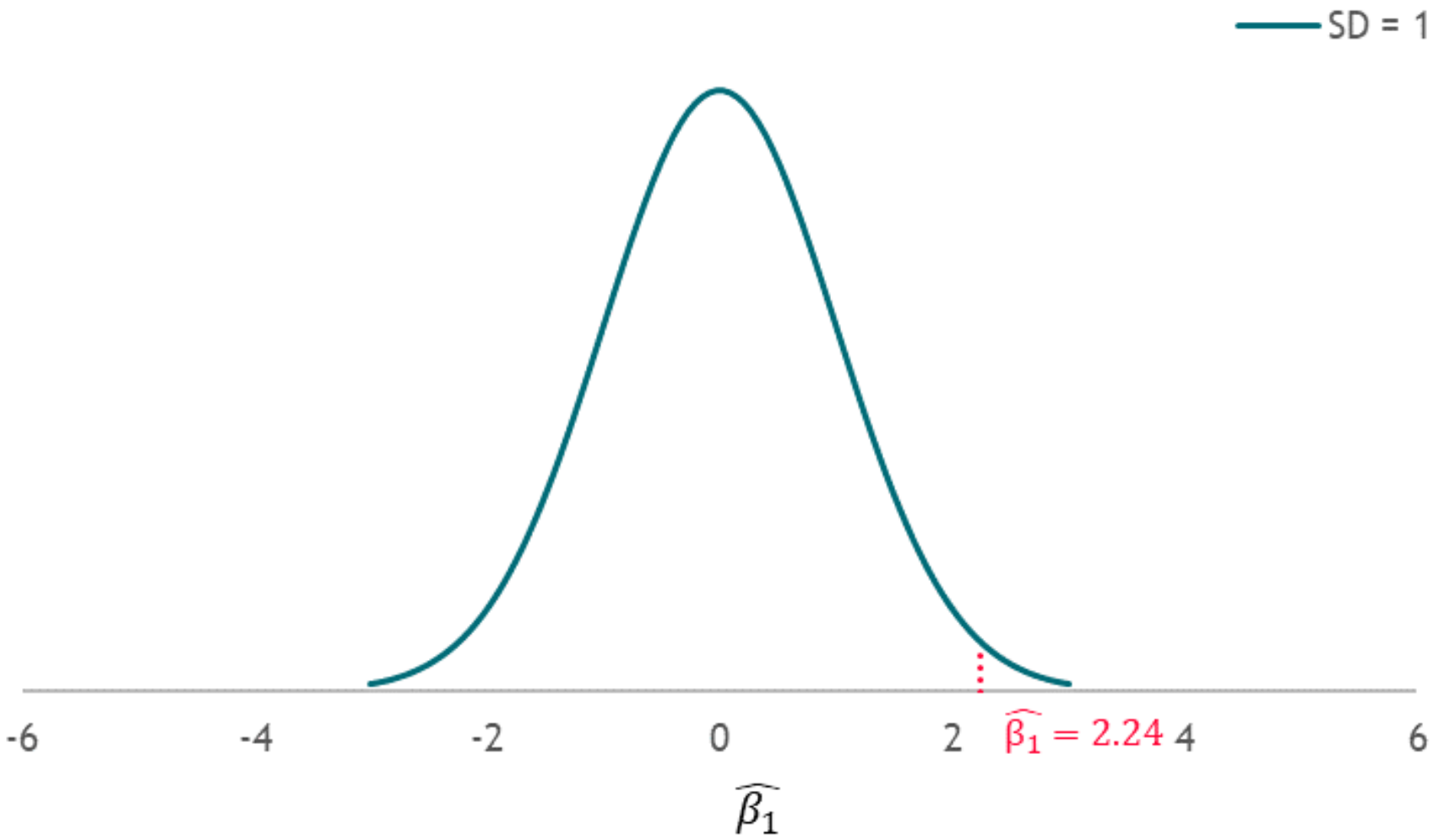
Let us test this

2. Second, we need to **estimate a statistic**.

Could we just use our $\hat{\beta}_1$ estimate and place it on the null distribution to see how unlikely it would be if the null were true?

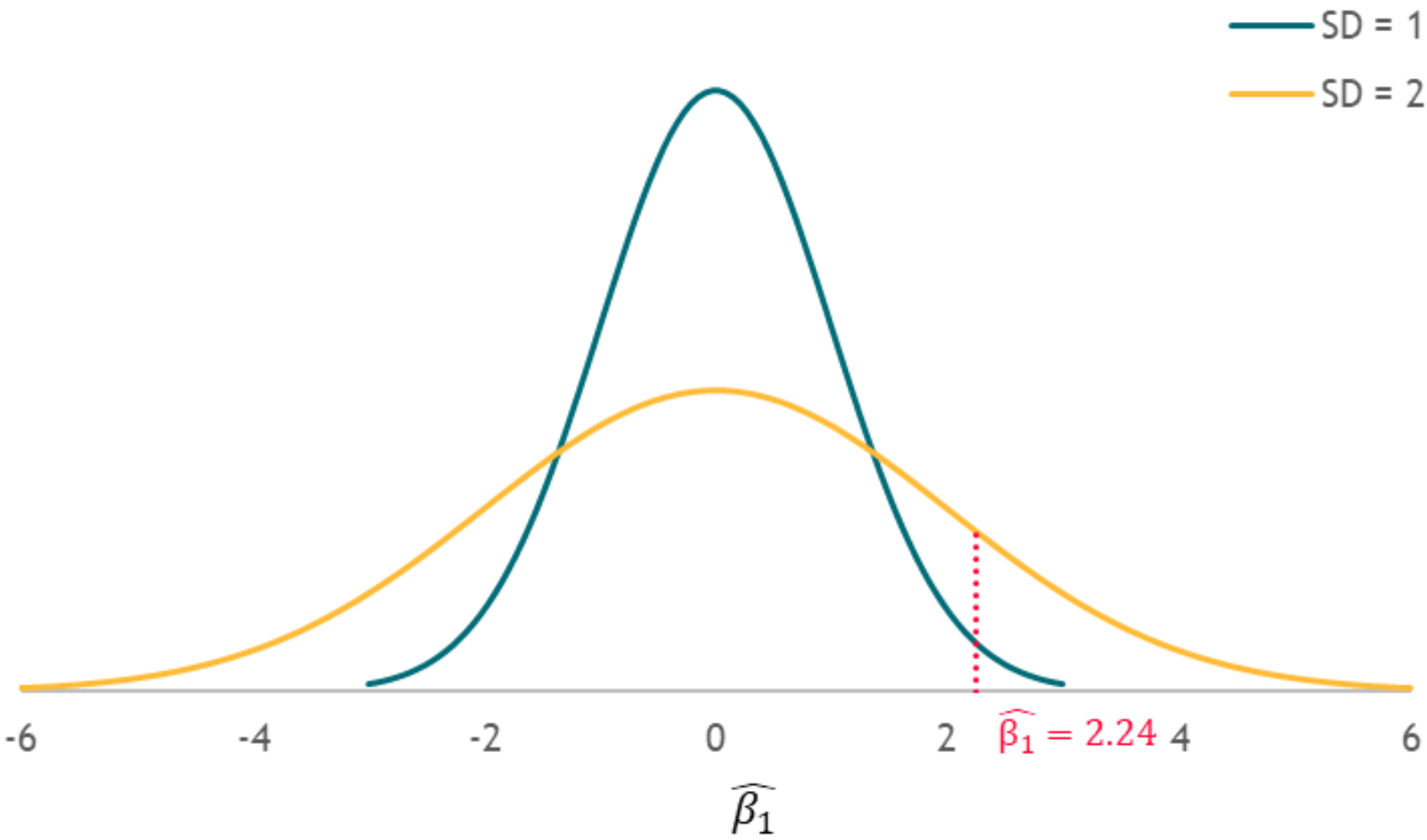
Problem: the scale of $\hat{\beta}_1$ could be anything!

Problem: where to place $\hat{\beta}_1$ depends on its spread



Here, the observed $\hat{\beta}_1$ is pretty *unlikely* if the null were true.

Problem: where to place $\hat{\beta}_1$ depends on its spread



When the variance is greater, the (same) observed $\hat{\beta}_1$ is *not so unlikely* if the null were true.

Solution: we standardize $\hat{\beta}_1$

Solution: we will standardize (rescale) $\hat{\beta}_1$ by dividing it by its standard deviation.

Pro-tip: remember that the standard error is an estimate of the standard deviation of a parameter.

This is the **t-statistic**!

$$\text{t-statistic} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

So we've added this bell and whistle to our machinery:

- ✓ How do we measure the outcome, i.e. what test statistic do we use?

What distribution should we use for the null?

3. Third, we need to **derive the null distribution**, i.e. the distribution of possible outcomes if the null hypothesis were true.

To do this, we need to answer the question: "what is the distribution of the t-statistic?".

Turns out, $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$ follows a t distribution.

(Thanks Guinness Brewing Company!)

The t distribution is *chunkier* than the normal distribution. It has fatter tails (that's a technical term!) than the normal when the sample size is small, which allows us to be more conservative.

t-distributions versus normal distributions

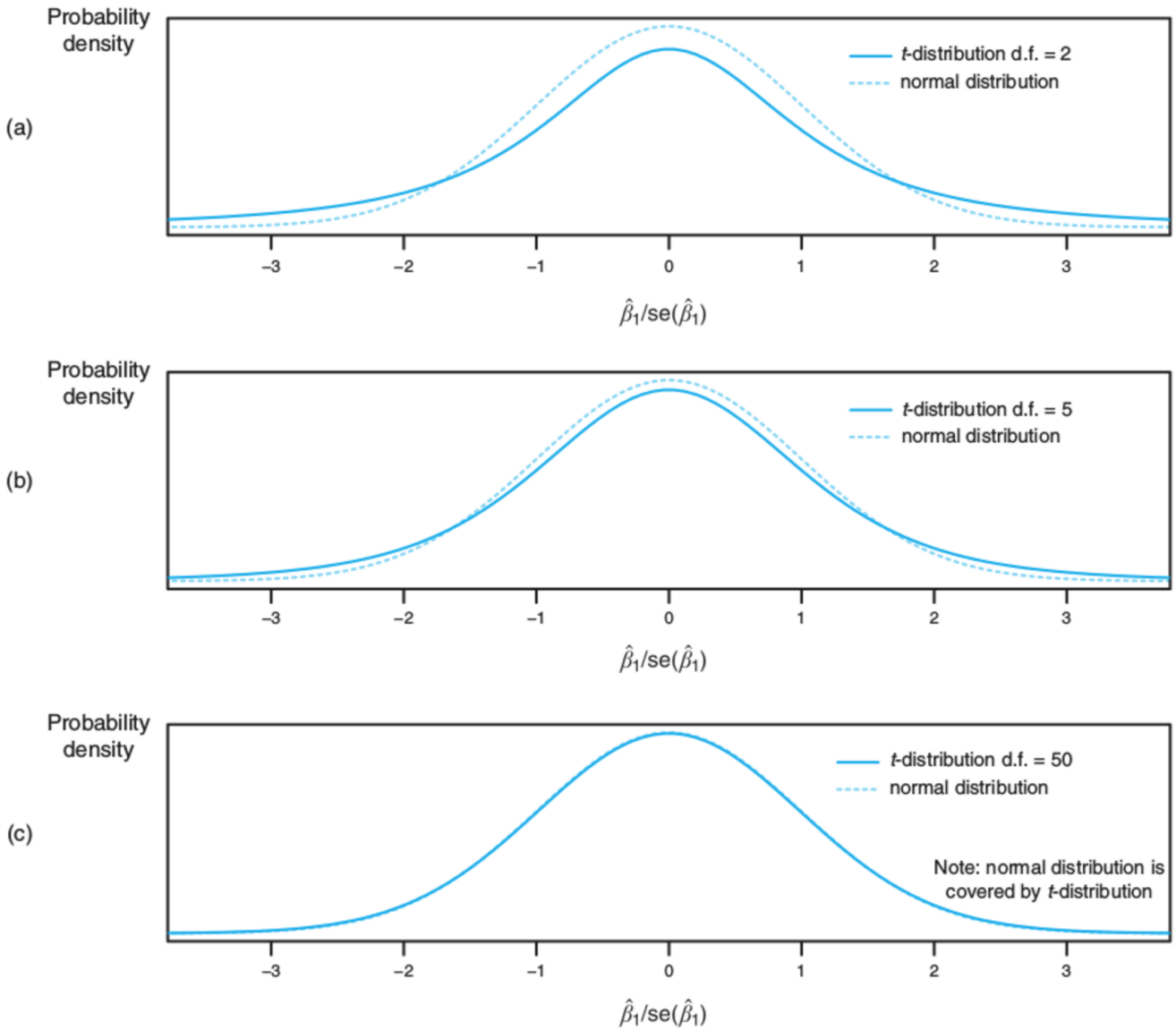


FIGURE 4.3: Three *t* Distributions

What distribution should we use for the null?

To test hypotheses in regression analysis, we use the **t-distribution** for the null hypothesis.

- Allows us to proceed with caution in small sample sizes (so that we do not incorrectly reject H_0).
- It is virtually indistinguishable from the normal distribution in large samples.

We've added the final piece to our machinery!

✓ What distribution do we use for the null? (E.g. normal, t)

Summary of hypothesis testing in regression analysis

1. **State your null hypothesis**
   - $H_0 : \beta_1 = 0$
2. Decide whether to use a **one-sided or two-sided alternative hypothesis**
   - $H_A : \beta_1 \neq 0$ (two-sided)
3. Decide on a **significance level** α
   - Conventionally, choose $\alpha = 0.05$
4. **Estimate** your regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
   - You obtain a $\hat{\beta}_1$ estimate
5. Compute your **test statistic** (i.e. standardize your coefficient)
   - t-statistic $= \hat{\beta}_1 / SE(\hat{\beta}_1)$
6. Decide **how unlikely your test statistic is under the null**. There are two (equivalent) ways to proceed here:
   1. p-value approach
   2. Critical value approach

Procedure for computing how "weird" result is

You want to decide how unlikely you would be to observe this effect if the null hypothesis were true. Both approaches will lead to the same conclusion.

1. **p-value approach**
   - Compute the p-value associated with your test statistic. (R does this for you).
   - Compare your computed p-value to your significance level α:
     - If p-value $< \alpha$, reject the null hypothesis H_0 in favor of the alternative H_A.
     - If p-value $> \alpha$, fail to reject the null hypothesis H_0.
2. **Critical value approach**
   - Look up the critical value c associated with your α level. Here, $c = 1.96$. Check. Why is that? Because $\alpha = 0.05$, we have chosen a two-sided test, and our sample is large.
   - Compare your computed t-statistic to your critical value c:
     - If |t-statistic| $> c$, reject the null hypothesis H_0 in favor of the alternative H_A.
     - If |t-statistic| $< c$, fail to reject the null hypothesis H_0.

Remember

	Reject the null H_0	**Fail to reject the null H_0**
p-value approach	p-value $< \alpha$	p-value $> \alpha$
Critical value approach	$\lvert$t-statistic$\rvert > c$	$\lvert$t-statistic$\rvert < c$

Note: for a two-sided H_A

- Significance level α and critical value c are thresholds to which you compare stuff to. You get to decide what these are (according to how cautious you want to be, more on that later).
- The p-value and the t-statistic are computed from your data. You don't get to decide what these are. You get what you get.

How to do hypothesis testing in R

Let's use the Fearon and Laitin data-set from Problem Sets 2 and 3.

Research question: Does ethnic fractionalization explain how long civil wars last?

- Dependent variable: years of civil war
- Independent variable: ethnic fractionalization

Hypothesis testing:

1. **State your null hypothesis**
   - $H_0 : \beta_1 = 0$ (there is no association between ethnic fractionalization and years of war)
2. **Decide on alternative hypothesis**
   - $H_A : \beta_1 \neq 0$ (there is *some* association between ethnic fractionalization and years of war)
3. Decide on a **significance level** α
   - Let's choose $\alpha = 0.05$

How to do hypothesis testing in R

4. **Estimate** your regression model
   - $numyears_i = \beta_0 + \beta_1 ethfrac_i + \epsilon_i$

```
load("fl2.RData")
model <- lm(numyears ~ ethfrac, data = fl2)
summary(model)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.571      2.079  21.439   <2e-16 ***
ethfrac       -9.830      4.205  -2.338   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We get $\hat{\beta}_1 = -9.830$

How to do hypothesis testing in R

5. Compute your **test statistic** (i.e. standardize your coefficient)

```
load("fl2.RData")
model <- lm(numyears ~ ethfrac, data = fl2)
summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.571      2.079  21.439   <2e-16 ***
ethfrac       -9.830      4.205  -2.338   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We get $SE(\hat{\beta}_1) = 4.205$

How to do hypothesis testing in R

5. Compute your **test statistic** (i.e. standardize your coefficient)

```
load("fl2.RData")
model <- lm(numyears ~ ethfrac, data = fl2)
summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.571      2.079  21.439   <2e-16 ***
ethfrac       -9.830      4.205  -2.338   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We get $SE(\hat{\beta}_1) = 4.205$
- t-statistic $= \hat{\beta}_1 / SE(\hat{\beta}_1) = -9.830/4.205 = -2.338$

How to do hypothesis testing in R

6. Decide **how unlikely your test statistic is under the null**.

```
load("fl2.RData")
model <- lm(numyears ~ ethfrac, data = fl2)
summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.571      2.079  21.439   <2e-16 ***
ethfrac       -9.830      4.205  -2.338   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Critical value approach
  - The critical value c corresponding to $\alpha = 0.05$ is 1.96
  - We find that $|-2.338| > 1.96$
  - We **reject the null hypothesis** that there is no association between ethnic fractionalization and years of civil war at $\alpha = 0.05$

How to do hypothesis testing in R

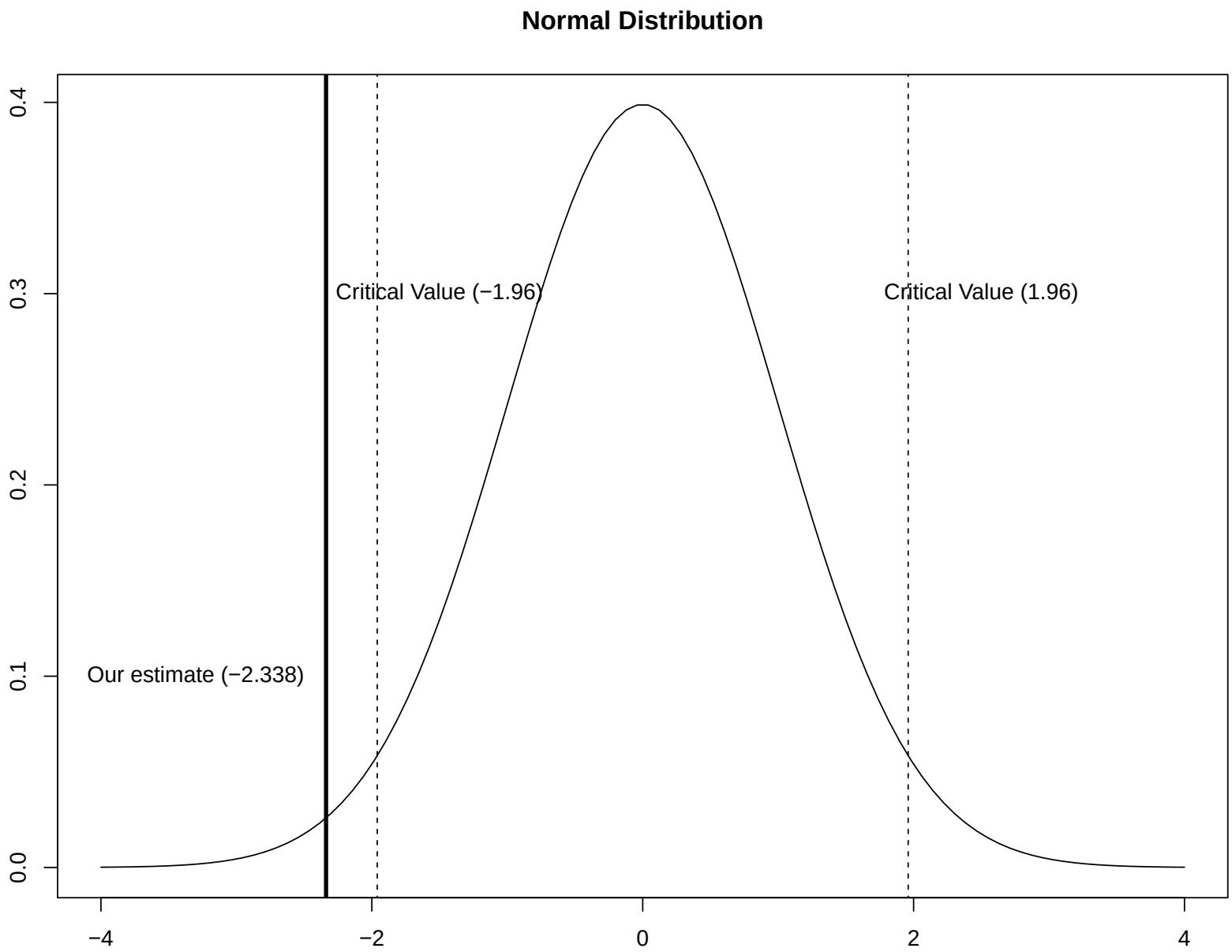
6. Decide **how unlikely your test statistic is under the null**.

```
load("fl2.RData")
model <- lm(numyears ~ ethfrac, data = fl2)
summary(model)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.571      2.079  21.439   <2e-16 ***
ethfrac       -9.830      4.205  -2.338   0.0207 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- p-value approach
  - We get a p-value of 0.0207
  - We find that $0.0207 < 0.05$
  - We **reject the null hypothesis** that there is no association between ethnic fractionalization and years of civil war at $\alpha = 0.05$

Visualizing Fearon and Laitin results



Political Science 15
Introduction to Research in Political Science
Lecture 5d: Type I and Type II errors

Alice Lépissier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Back to the big picture

Remember, we never prove or disprove the null hypothesis.

When reject H_0, we are saying that, given the effect we observe, it is *unlikely* that $\beta_1 = 0$, but not *impossible*.

Thinking probabilistically means recognizing that we might make a mistake.

Let's now talk about what types of mistakes we can make and how costly they can be.

Two types of mistakes

False positive (Type I error)

- When we reject a null hypothesis that is actually true: **we say there is a relationship when there isn't one**.
- Recall that the definition of α is how unlikely our result $\hat{\beta}_1$ has to be under the null, for us to be able to reject the null.
- So if we set $\alpha = 0.05$, we still have a 5% chance that the $\hat{\beta}_1$ we observed is high enough that we reject H_0 even when it is true.
- In which case we would be making a type I error (false positive)!

Our amazing significance level α

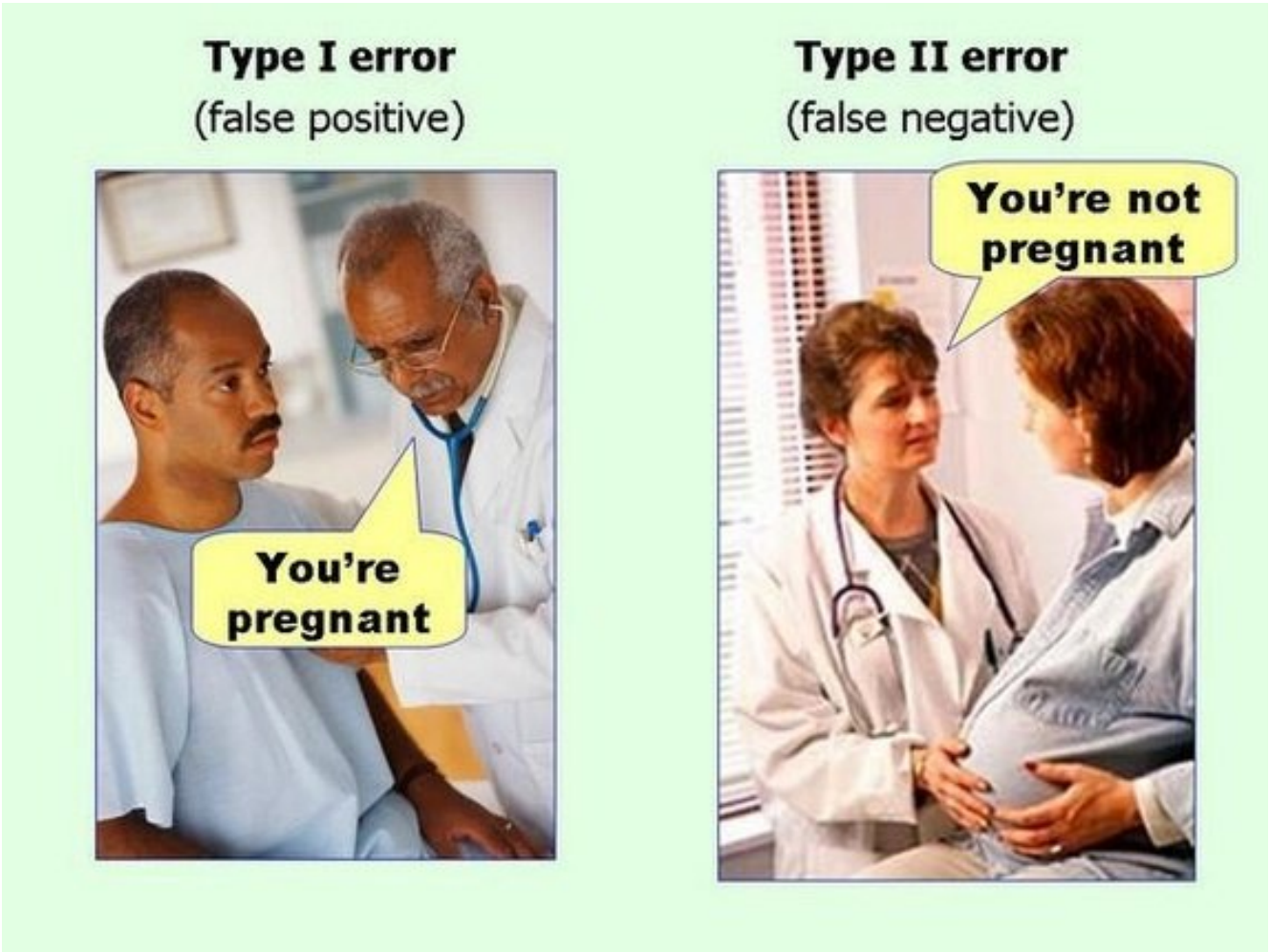
- The significance level α is the **probability of committing a type I error**!
- Much of statistics is concerned with limiting this type of error.
- This is why we choose conservative levels of α: to minimize the probability of False Positives.

Two types of mistakes

False negative (Type II error)

- When we fail to reject the null, but the null is actually false: **we say there is no relationship, when there is one**.
- This could happen if we have a small sample size (the study has low power).
  - A low sample size tends to increase the standard errors, so our null distribution is quite wide $\rightarrow$ harder to reject the null.
- The hypothesis testing framework is concerned with limiting the rate of False Positives.
- But there is a **trade-off** between the rate of False Positives and the rate of False Negatives.
  - You can reduce your risk of False Positives by using a lower value for α, e.g., setting $\alpha = 0.01$ means there is a 1% chance of committing a type I error.
  - But, using a lower value for α means that you will be less likely to detect a true difference if one really exists (thus risking a type II error).

		Reality	
		Null is False	**Null is True**
Decision	**Reject** H_0	Correct	Type I Error (FP)
	Don't reject H_0	Type II Error (FN)	Correct



Statistical hypothesis testing: the trial analogy

The decision theory that underlies statistical hypothesis testing is similar to a judicial trial.

Suppose we must decide whether to convict or acquit a defendant based on evidence presented at a trial. There are four possible outcomes.

		State of the World	
		Guilty	**Innocent**
Decision	**Convict**	Correct	Type I Error
	Acquit	Type II Error	Correct

Our goal is to limit the probability of error.

Our null hypothesis is H_0: the defendant is innocent.
H_0 is presumed to be true unless the data/evidence strongly suggest otherwise and we may be willing to reject the null hypothesis in favor of an alternative hypothesis.

Statistical hypothesis testing: the trial analogy

Suppose we can somehow model the probabilities for the various outcomes conditional on the true state of the world.

Probabilities given the true state of the world

		State of the World	
		Guilty	**Innocent**
Decision	**Convict**	$1-\beta$	α
	Acquit	β	$1-\alpha$

We would like α and β to be small, but it is difficult to achieve both goals at the same time. Open question: which type of mistake is worst?

- **Type I error example**. A defendant is accused of a heinous crime. They are sentenced to life in prison, when in fact they are innocent.
- **Type II error example**. The result of a patient's biopsy is negative, when in fact their tumor is malign. They do not seek treatment for cancer, and their life is shortened as a result.

Bottom line

The consequences of making a Type I error (False Positive) or a Type II error (False Negative) depends on the underlying question we are trying to answer.

Balancing the trade-off between the rate of False Positives and the rate of False Negatives is salient in many different fields, e.g. judicial trial, machine learning, disease testing.

The statistical hypothesis testing framework (AKA Lecture 5) tends to be concerned with limiting the rate of False Positives (by controlling the significance level α).

You can limit the rate of False Negatives through research design: by increasing your sample size, you increase your study's statistical power.