

# Simple Linear Regression: Inference

EC 320: Introduction to Econometrics

Kyle Raze

Fall 2019

# Prologue

# Housekeeping

## Problem Set 3

- Due tonight by 11:59pm.

## Problem Set 4

- Due Monday, November 18 by 11:59pm.
- 10 points extra credit for using R Markdown to generate responses to the analytical problems (HTML or Word).

# Last Time

We discussed the **classical assumptions of OLS**:

1. **Linearity:** The population relationship is linear in parameters with an additive error term.
2. **Sample Variation:** There is variation in  $X$ .
3. **Random Sampling:** We have a random sample from the population of interest.
4. **Exogeneity:** The  $X$  variable is exogenous (*i.e.*,  $\mathbb{E}(u|X) = 0$ ).
5. **Homoskedasticity:** The error term has the same variance for each value of the independent variable (*i.e.*,  $\text{Var}(u|X) = \sigma^2$ ).
6. **Normality:** The population error term is normally distributed with mean zero and variance  $\sigma^2$  (*i.e.*,  $u \sim N(0, \sigma^2)$ )

We restricted our attention to the first 5 assumptions.

# Classical Assumptions

## Last Time

1. We used the first 4 assumptions to show that OLS is unbiased:

$$\mathbb{E}[\hat{\beta}] = \beta$$

2. We used the first 5 assumptions to derive a formula for the **variance** of the OLS estimator:  $\text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$ .

# Classical Assumptions

## Today

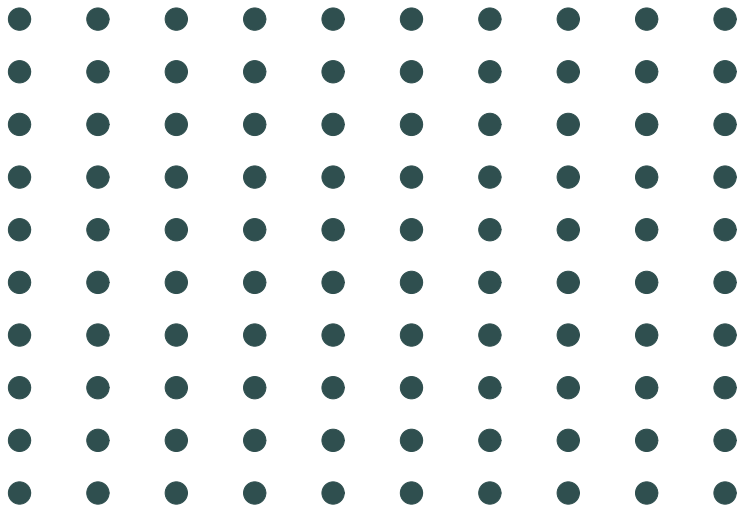
We will use the sampling distribution of  $\hat{\beta}$  to conduct hypothesis tests.

- Can use all 6 classical assumptions to show that OLS is normally distributed:

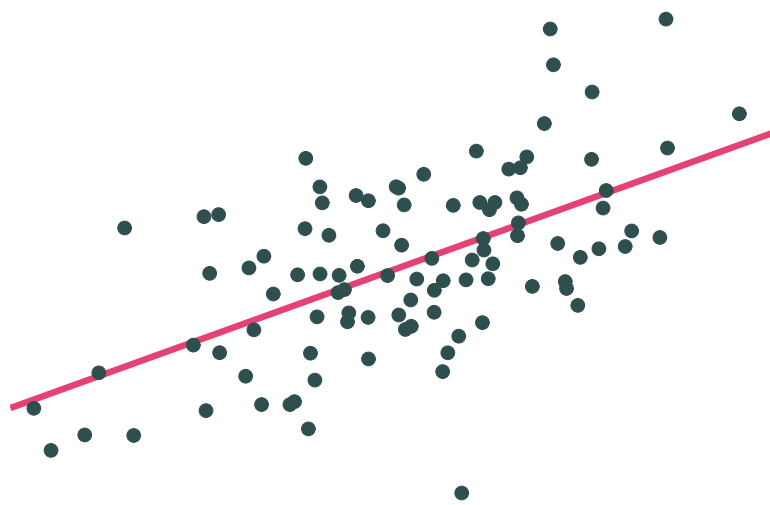
$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

- We'll "prove" this using R.

# Simulation



**Population**

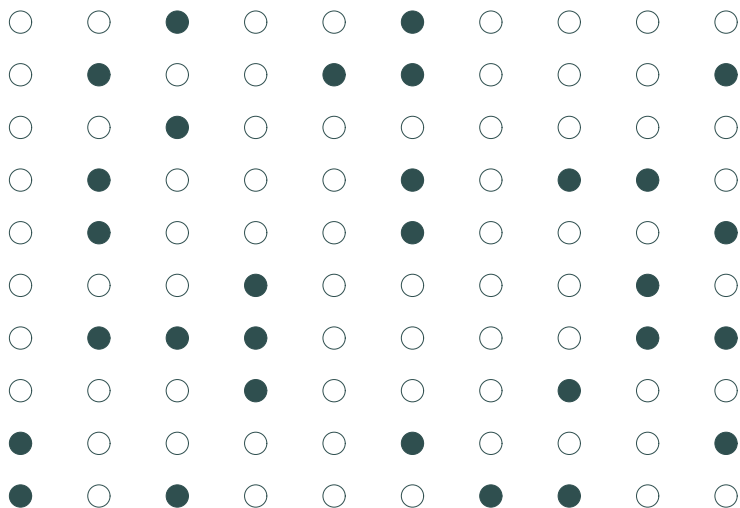


**Population relationship**

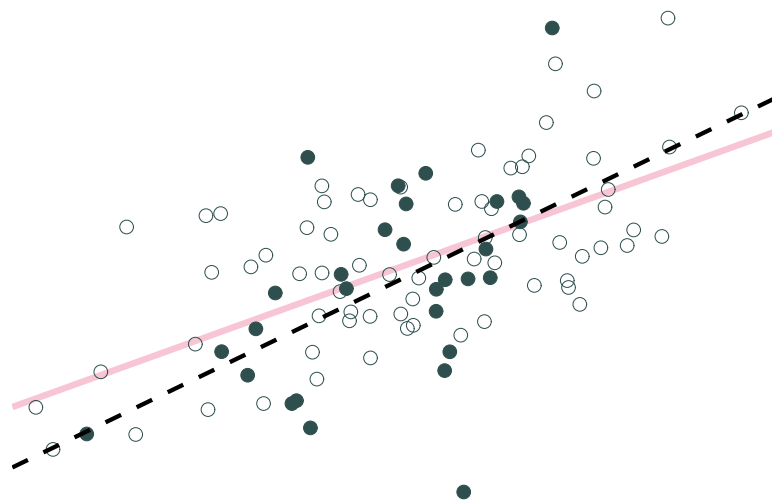
$$Y_i = 2.53 + 0.57X_i + u_i$$

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

# Simulation



**Sample 1:** 30 random individuals



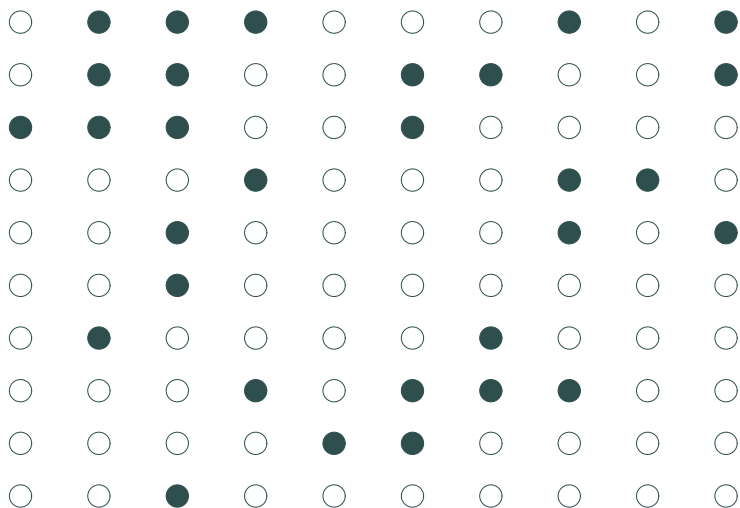
**Population relationship**

$$Y_i = 2.53 + 0.57X_i + u_i$$

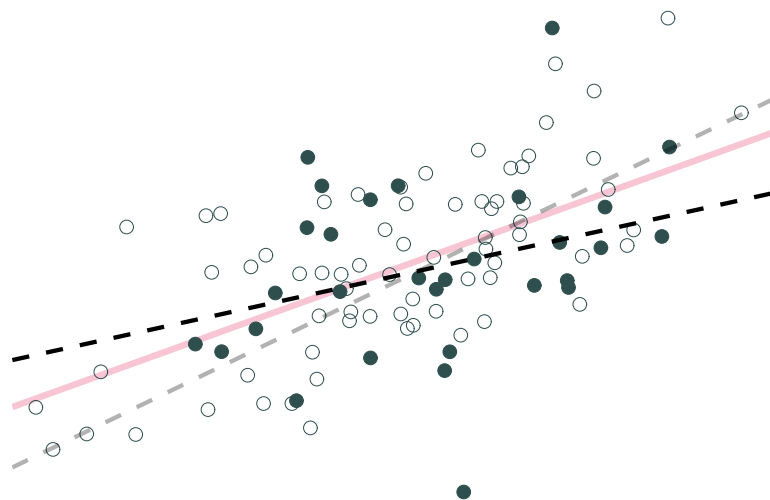
**Sample relationship**

$$\hat{Y}_i = 1.36 + 0.76X_i$$

# Simulation



**Sample 2:** 30 random individuals



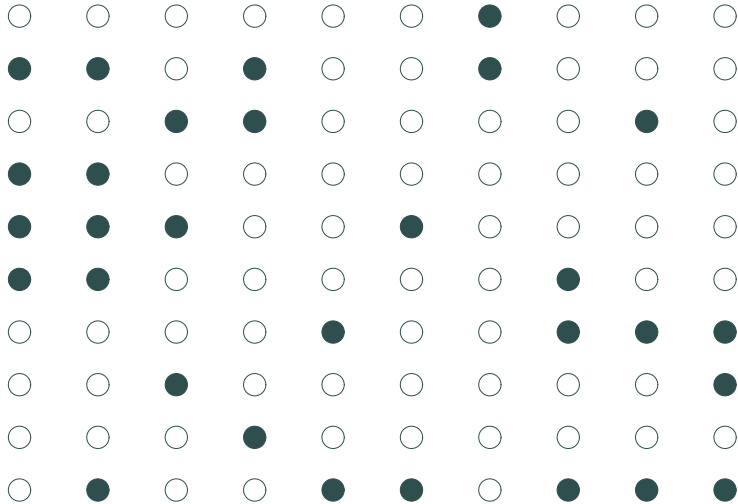
**Population relationship**

$$Y_i = 2.53 + 0.57X_i + u_i$$

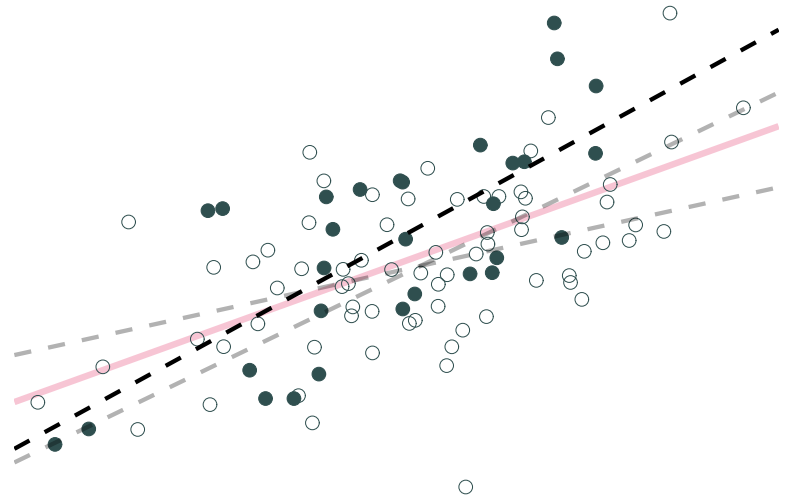
**Sample relationship**

$$\hat{Y}_i = 3.53 + 0.34X_i$$

# Simulation



**Sample 3:** 30 random individuals



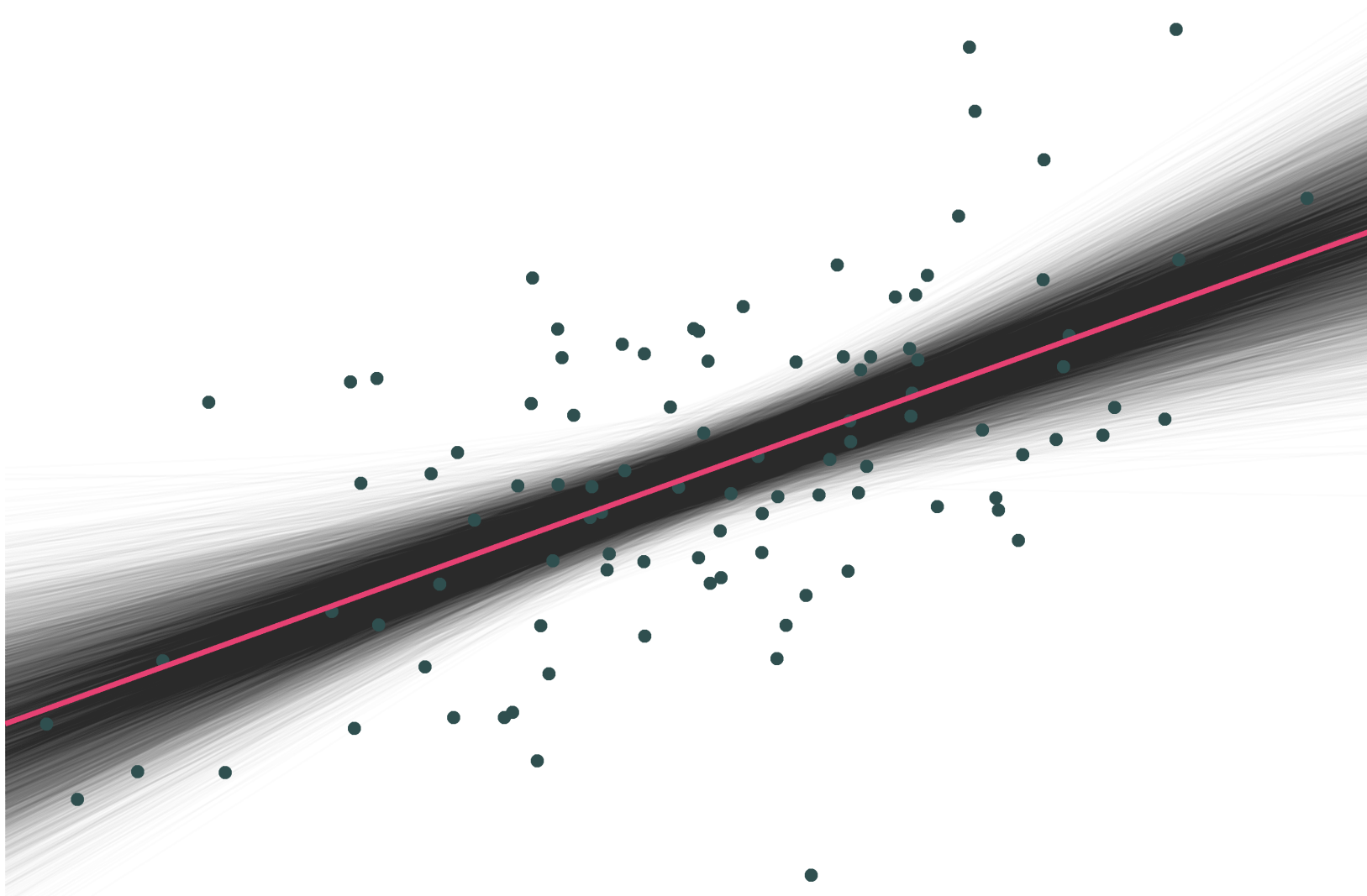
**Population relationship**

$$Y_i = 2.53 + 0.57X_i + u_i$$

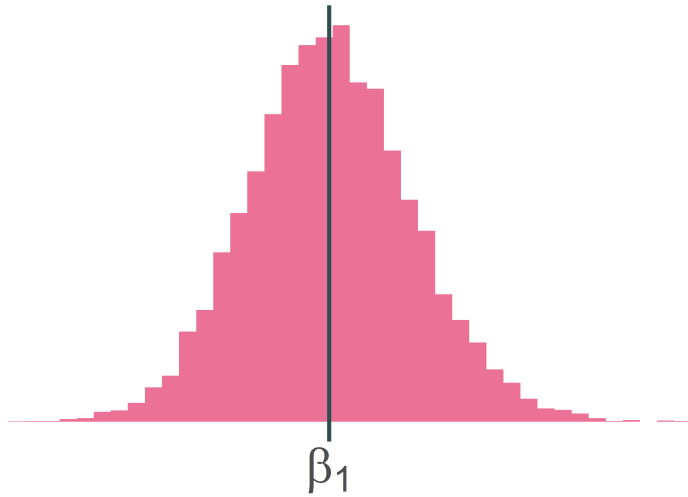
**Sample relationship**

$$\hat{Y}_i = 1.44 + 0.86X_i$$

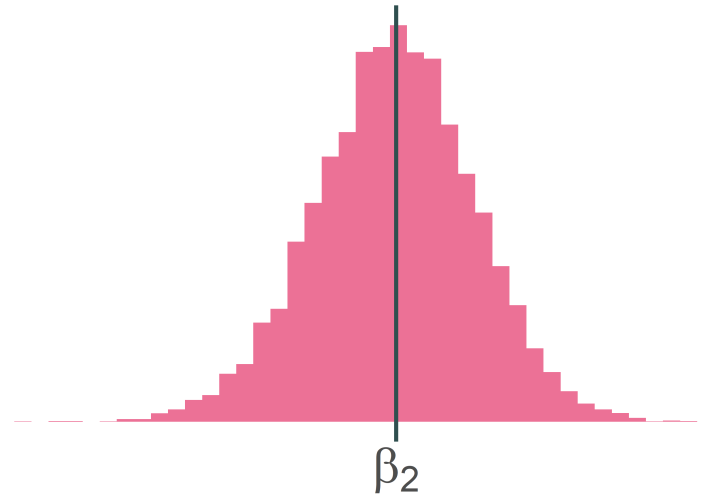
Repeat **10,000 times** (Monte Carlo simulation).



## Intercept Estimates



## Slope Estimates



# Simulation

Can you spot the classical assumptions?

```
# Set population and sample sizes
n_p ← 100
n_s ← 30
# Generate population data
pop_df ← tibble(
  x = rnorm(n_p, mean = 5, sd = 1.5),
  e = rnorm(n_p, mean = 0, sd = 1),
  y = 2.53 + 0.57 * x + e
)
# Define simulation procedure
sim_ols ← function(x, size = n_s) {
  lm(y ~ x, data = pop_df %>% sample_n(size = size)) %>%
    tidy() %>%
    mutate(iteration = x)
}
# Run simulation
sim_df ← map_df(1:10000, ~sim_ols(.x, size = n_s))
```

# Inference

# Motivation

What does statistical evidence say about existing theories?

We want to test hypotheses posed by politicians, economists, scientists, people with foil hats, *etc.*

- Does building a giant wall **reduce crime**?
- Does shutting down a government **adversely affect the economy**?
- Does legal cannabis **reduce drunk driving** or **reduce opioid use**?
- Do air quality standards **improve health** or **reduce jobs**?

While uncertainty exists, we can still conduct *reliable* statistical tests (rejecting or failing to reject a hypothesis).

# Inference

We know OLS has some nice properties, and we know how to estimate an intercept and slope coefficient using OLS.

Our current workflow:

- Get data (points with  $X$  and  $Y$  values).
- Regress  $Y$  on  $X$ .
- Plot the fitted values (i.e.,  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ) and report the estimates.

But how do we actually **learn** something from this exercise?

- Based upon our value of  $\hat{\beta}_2$ , can we rule out previously hypothesized values?
- How confident should we be in the precision of our estimates?

We need to be able to deal with uncertainty. Enter: **Inference**.

# Inference

We use the standard error of  $\hat{\beta}_2$ , along with  $\hat{\beta}_2$  itself, to learn about the parameter  $\beta_2$ .

After deriving the distribution of  $\hat{\beta}_2$ ,<sup>†</sup> we have two (related) options for formal statistical inference (learning) about our unknown parameter  $\beta_2$ :

- **Hypothesis tests:** Determine whether there is statistically significant evidence to reject a hypothesized value or range of values.
- **Confidence intervals:** Use the estimate and its standard error to create an interval that, when repeated, will generally<sup>††</sup> contain the true parameter.

<sup>†</sup> Hint: It's normal with mean  $\beta_2$  and variance  $\frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$ .

<sup>††</sup> E.g., similarly constructed 95% confidence intervals will contain the true parameter 95% of the time.

# OLS Variance

Hypothesis tests and confidence intervals require information about the variance of the OLS estimator:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

## Problem

- The variance formula has a population parameter:  $\sigma^2$  (a.k.a. error variance).
- We can't observe population parameters.
- **Solution:** Estimate  $\sigma^2$ .

# Estimating Error Variance

## Learning from our (prediction) errors

We can estimate the variance of  $u_i$  (a.k.a.  $\sigma^2$ ) using the sum of squared residuals:

$$s_u^2 = \frac{\sum_i \hat{u}_i^2}{n - k}$$

where  $k$  gives the number of regression parameters.

- In a simple linear regression,  $k = 2$ .
- $s_u^2$  is an unbiased estimator of  $\sigma^2$ .

# OLS Variance, Take 2

With  $s_u^2 = \frac{\sum_i \hat{u}_i^2}{n - k}$ , we can calculate

$$\text{Var}(\hat{\beta}_2) = \frac{s_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Taking the square root, we get the **standard error** of the OLS estimator:

$$\text{SE}(\hat{\beta}_2) = \sqrt{\frac{s_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}.$$

- Standard error = standard deviation of an estimator.

# Standard Errors

R's `lm()` function estimates standard errors out of the box:

```
tidy(lm(y ~ x, pop_df))
```

```
#> # A tibble: 2 x 5
#>   term          estimate std.error statistic  p.value
#>   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
#> 1 (Intercept)    2.53      0.422      6.00 3.38e- 8
#> 2 x              0.567     0.0793     7.15 1.59e-10
```

I won't ask you to estimate standard errors by hand!

# Hypothesis Tests

# Hypothesis Tests

**Null hypothesis ( $H_0$ ):**  $\beta_2 = 0$

**Alternative hypothesis ( $H_a$ ):**  $\beta_2 \neq 0$

There are four possible outcomes of our test:

1. We **fail to reject** the null hypothesis and the null is true.
2. We **reject** the null hypothesis and the null is false.
3. We **reject** the null hypothesis, but the null is actually true (**Type I error**).
4. We **fail to reject** the null hypothesis, but the null is actually false (**Type II error**).

# Hypothesis Tests

**Goal:** Make a statement about  $\beta_2$  using information on  $\hat{\beta}_2$ .

$\hat{\beta}_2$  is random: it could be anything, even if  $\beta_2 = 0$  is true.

- But if  $\beta_2 = 0$  is true, then  $\hat{\beta}_2$  is unlikely to take values far from zero.
- As the standard error shrinks, we are even less likely to observe "extreme" values of  $\hat{\beta}_2$  (assuming  $\beta_2 = 0$ ).

Our test should take **extreme values** of  $\hat{\beta}_2$  as **evidence against the null hypothesis**, but it should also weight them by what we know about the variance of  $\hat{\beta}_2$ .

# Hypothesis Tests

## Null hypothesis

$$H_0: \beta_2 = 0$$

## Alternative hypothesis

$$H_a: \beta_2 \neq 0$$

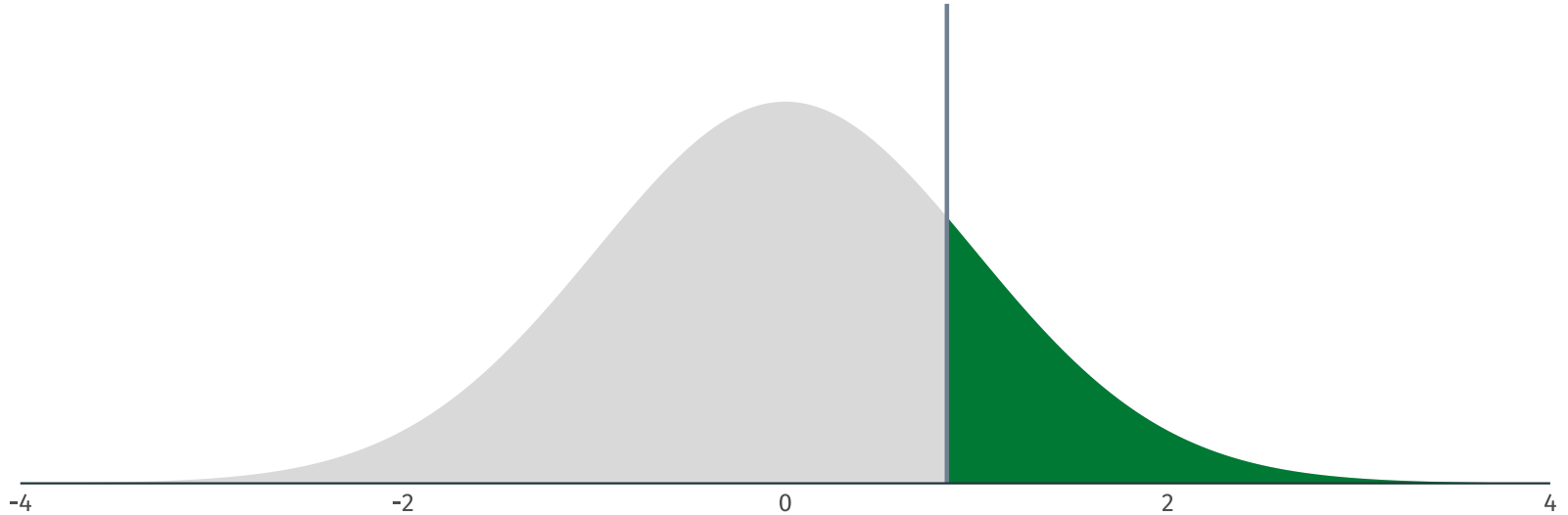
To conduct the test, we calculate a  $t$ -statistic:

$$t = \frac{\hat{\beta}_2 - \beta_2^0}{\text{SE}(\hat{\beta}_2)}$$

- Distributed according to a  $t$ -distribution with  $n - 2$  *degrees of freedom*.
- $\beta_2^0$  is the value of  $\beta_2$  in our null hypothesis (e.g.,  $\beta_2^0 = 0$ ).

# Hypothesis Tests

Next, we use the  $t$ -**statistic** to calculate a  $p$ -**value**.



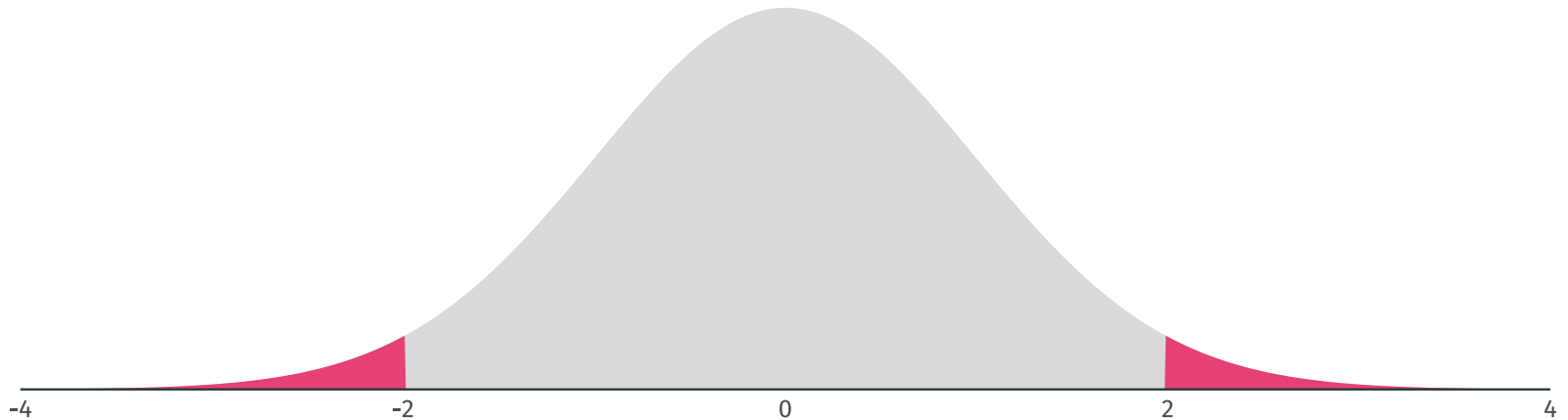
Describes the probability of seeing a  $t$ -statistic as extreme as the one we observe *if the null hypothesis is actually true*.

But...we still need some benchmark to compare our  $p$ -value against.

# Hypothesis Tests

We worry mostly about false positives, so we conduct hypothesis tests based on the probability of making a Type I error.

**How?** We select a **significance level  $\alpha$**  that specifies our tolerance for false positives. This is the probability of Type I error we choose to live with.



# Hypothesis Tests

We then compare  $\alpha$  to the  $p$ -value of our test.

- If the  $p$ -value is less than  $\alpha$ , then we **reject the null hypothesis** at the  $\alpha \cdot 100$  percent level.
- If the  $p$ -value is greater than  $\alpha$ , then we **fail to reject the null hypothesis**.
- **Note:** *Fail to reject  $\neq$  accept.*

# Hypothesis Tests

**Example:** Are campus police associated with campus crime?

```
lm(crime ~ police, data = campus) %>% tidy()
```

```
#> # A tibble: 2 x 5  
#>   term          estimate std.error statistic  p.value  
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>  
#> 1 (Intercept)    18.4       2.38      7.75 1.06e-11  
#> 2 police         1.76       1.30      1.35 1.81e- 1
```

$H_0: \beta_{\text{Police}} = 0$  v.s.  $H_a: \beta_{\text{Police}} \neq 0$

Significance level:  $\alpha = 0.05$  (i.e., 5 percent test)

Test Condition: Reject  $H_0$  if  $p < \alpha$

$p = 0.18$ . **Do we reject the null hypothesis?**

# Hypothesis Tests

$p$ -values are difficult to calculate by hand.

**Alternative:** Compare  $t$ -statistic to critical values from the  $t$ -distribution.



# Hypothesis Tests

**Notation:**  $t_{1-\alpha/2, n-2}$  or  $t_{\text{crit}}$ .

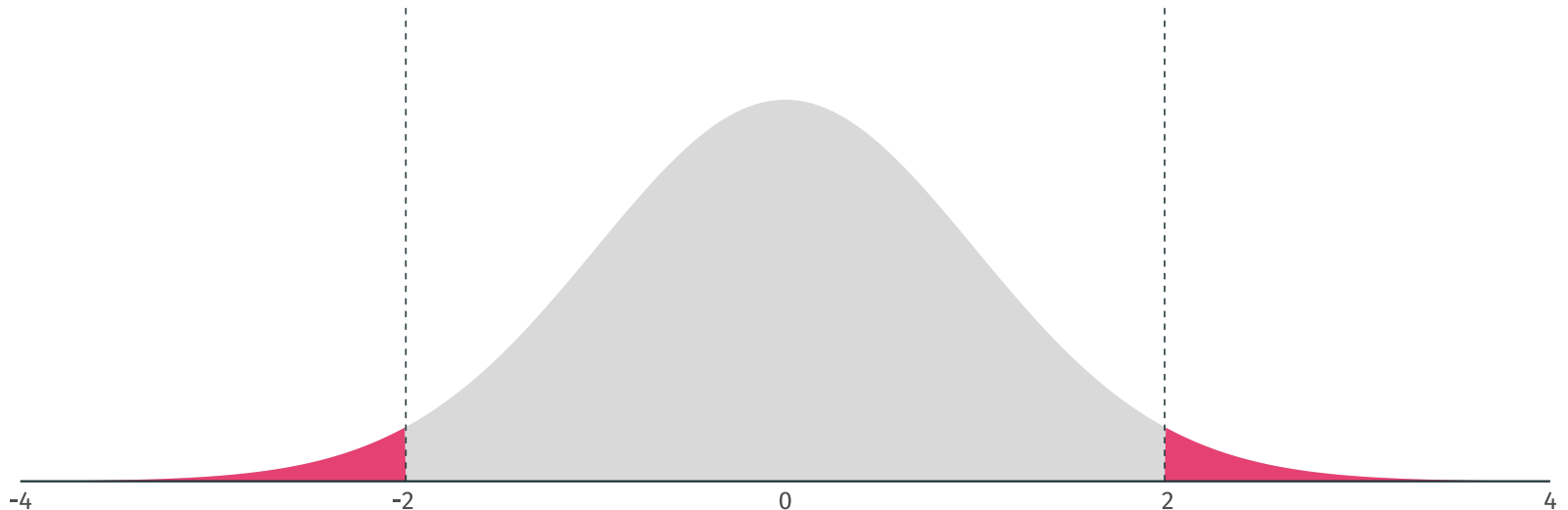
- Find in a  $t$  table using the significance level  $\alpha$  and  $n - 2$  degrees of freedom.

Compare the the critical value to your  $t$ -statistic:

- If  $|t| > |t_{1-\alpha/2, n-2}|$ , then **reject the null**.
- If  $|t| < |t_{1-\alpha/2, n-2}|$ , then **fail to reject the null**.

# Two-Sided Tests

Based on a critical value of  $t_{1-\alpha/2, n-2} = t_{0.975, 100} = 1.98$ , we can identify a **rejection region** on the  $t$ -distribution.



If our  $t$  statistic is in the rejection region, then we reject the null hypothesis at the 5 percent level.

# Two-Sided Tests

R defaults to testing hypotheses against the null hypothesis of zero.

```
lm(y ~ x, data = pop_df) %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term          estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    2.53      0.422      6.00 3.38e- 8
#> 2 x              0.567     0.0793     7.15 1.59e-10
```

$H_0: \beta_2 = 0$  vs.  $H_a: \beta_2 \neq 0$

Significance level:  $\alpha = 0.05$  (*i.e.*, 5 percent test)

$t_{\text{stat}} = 7.15$  and  $t_{0.975, 28} = 2.05$ , which implies that  $p < 0.05$ .

Therefore, we **reject  $H_0$**  at the 5% level.

# Two-Sided Tests

**Example:** Are campus police associated with campus crime?

```
lm(crime ~ police, data = campus) %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term          estimate std.error statistic  p.value
#>   <chr>          <dbl>    <dbl>    <dbl>    <dbl>
#> 1 (Intercept)    18.4      2.38     7.75 1.06e-11
#> 2 police         1.76      1.30     1.35 1.81e- 1
```

$H_0: \beta_{\text{Police}} = 0$  v.s.  $H_a: \beta_{\text{Police}} \neq 0$

Significance level:  $\alpha = 0.1$  (i.e., 10 percent test)

Test Condition: Reject  $H_0$  if  $|t| > t_{\text{crit}}$

$t = 1.35$  and  $t_{\text{crit}} = 1.66$ . **Do we reject the null hypothesis?**

# One-Sided Tests

Sometimes we are confident that a parameter is non-negative or non-positive.

A **one-sided** test assumes that values on one side of the null hypothesis are impossible.

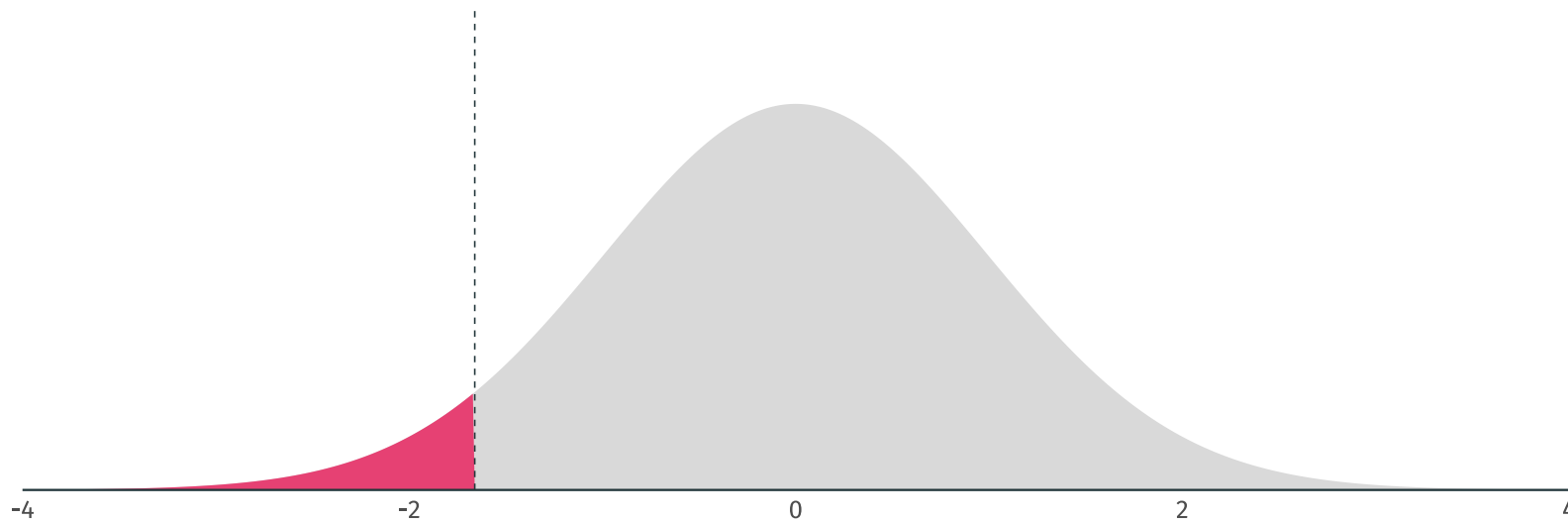
- **Option 1:**  $H_0: \beta_2 = 0$  vs.  $H_a: \beta_2 > 0$
- **Option 2:**  $H_0: \beta_2 = 0$  vs.  $H_a: \beta_2 < 0$

If this assumption is reasonable, then our rejection region changes.

- Same  $\alpha$ .

# One-Sided Tests

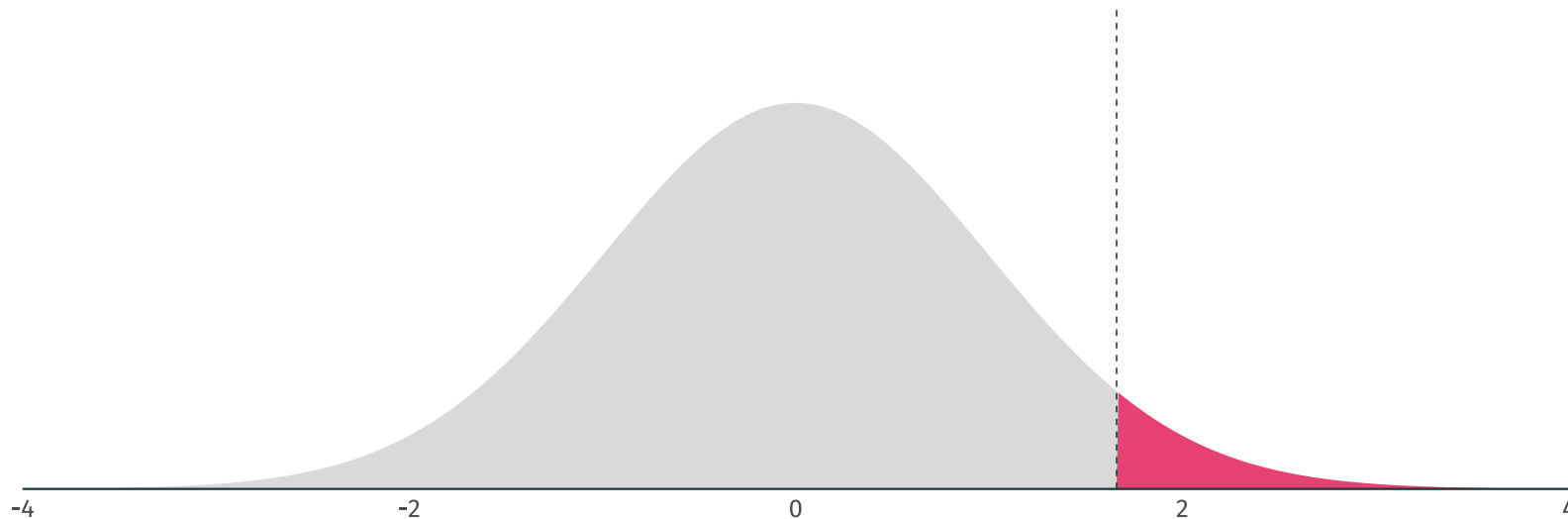
**Left-tailed:** Based on a critical value of  $t_{1-\alpha, n-2} = t_{0.95, 100} = 1.66$ , we can identify a **rejection region** on the  $t$ -distribution.



If our  $t$  statistic is in the rejection region, then we reject the null hypothesis at the 5 percent level.

# One-Sided Tests

**Right-tailed:** Based on a critical value of  $t_{1-\alpha, n-2} = t_{0.95, 100} = 1.66$ , we can identify a **rejection region** on the  $t$ -distribution.



If our  $t$  statistic is in the rejection region, then we reject the null hypothesis at the 5 percent level.

# One-Sided Tests

**Example:** Do campus police deter campus crime?

```
lm(crime ~ police, data = campus) %>% tidy()
```

```
#> # A tibble: 2 x 5  
#>   term          estimate std.error statistic  p.value  
#>   <chr>          <dbl>    <dbl>    <dbl>    <dbl>  
#> 1 (Intercept)    18.4      2.38     7.75 1.06e-11  
#> 2 police         1.76      1.30     1.35 1.81e- 1
```

$H_0: \beta_{\text{Police}} = 0$  v.s.  $H_a: \beta_{\text{Police}} < 0$

Significance level:  $\alpha = 0.1$  (i.e., 10 percent test)

Test Condition: Reject  $H_0$  if  $t < -t_{\text{crit}}$

$t = 1.35$  and  $t_{\text{crit}} = 1.29$ . **Do we reject the null hypothesis?**

# Confidence Intervals

# Confidence Intervals

Until now, we have considered **point estimates** of population parameters.

- Sometimes a range of values is more interesting/honest.

We can construct  $(1 - \alpha) \cdot 100$ -percent level confidence intervals for  $\beta_2$

$$\hat{\beta}_2 \pm t_{1-\alpha/2, n-2} \text{SE}(\hat{\beta}_2)$$

$t_{1-\alpha/2, n-2}$  denotes the  $1 - \alpha/2$  quantile of a  $t$  distribution with  $n - 2$  degrees of freedom.

# Confidence Intervals

**Q:** Where does the confidence interval formula come from?

**A:** The confidence interval formula comes from the rejection condition of a two-sided test.

Reject  $H_0$  if  $|t| > t_{\text{crit}}$

The test condition implies

Fail to reject  $H_0$  if  $|t| \leq t_{\text{crit}}$

which is equivalent to

Fail to reject  $H_0$  if  $-t_{\text{crit}} \leq t \leq t_{\text{crit}}$ .

# Confidence Intervals

Replacing  $t$  with its formula gives

$$\text{Fail to reject } H_0 \text{ if } -t_{\text{crit}} \leq \frac{\hat{\beta}_2 - \beta_2^0}{\text{SE}(\hat{\beta}_2)} \leq t_{\text{crit}}.$$

Standard errors are always positive, so the inequalities do not flip when we multiply by  $\text{SE}(\hat{\beta}_2)$ :

$$\text{Fail to reject } H_0 \text{ if } -t_{\text{crit}} \text{SE}(\hat{\beta}_2) \leq \hat{\beta}_2 - \beta_2^0 \leq t_{\text{crit}} \text{SE}(\hat{\beta}_2).$$

Subtracting  $\hat{\beta}_2$  yields

$$\text{Fail to reject } H_0 \text{ if } -\hat{\beta}_2 - t_{\text{crit}} \text{SE}(\hat{\beta}_2) \leq -\beta_2^0 \leq -\hat{\beta}_2 + t_{\text{crit}} \text{SE}(\hat{\beta}_2).$$

# Confidence Intervals

Multiplying by -1 and rearranging gives

Fail to reject  $H_0$  if

$$\hat{\beta}_2 - t_{\text{crit}} \hat{\text{SE}}(\hat{\beta}_2) \leq \beta_2^0 \leq \hat{\beta}_2 + t_{\text{crit}} \hat{\text{SE}}(\hat{\beta}_2).$$

Replacing  $\beta_2^0$  with  $\beta_2$  and dropping the test condition yields the interval

$$\hat{\beta}_2 - t_{\text{crit}} \hat{\text{SE}}(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\text{crit}} \hat{\text{SE}}(\hat{\beta}_2)$$

which is equivalent to

$$\hat{\beta}_2 \pm t_{\text{crit}} \hat{\text{SE}}(\hat{\beta}_2).$$

# Confidence Intervals

**Insight:** A confidence interval is related to a two-sided hypothesis test.

- If a 95 percent confidence interval contains zero, then we fail to reject the null hypothesis at the 5 percent level.
- If a 95 percent confidence interval does not contain zero, then we reject the null hypothesis at the 5 percent level.
- **Generally:** A  $(1 - \alpha) \cdot 100$  percent confidence interval embeds a two-sided test at the  $\alpha \cdot 100$  level.

# Confidence Intervals

## Example

```
lm(y ~ x, data = pop_df) %>% tidy()
```

```
#> # A tibble: 2 x 5  
#>   term          estimate std.error statistic  p.value  
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>  
#> 1 (Intercept)    2.53      0.422      6.00 3.38e- 8  
#> 2 x              0.567     0.0793     7.15 1.59e-10
```

```
# find degrees of freedom  
dof ← summary(lm(y ~ x, data = pop_df))$df[2]  
# return critical value  
qt(0.975, dof)
```

```
#> [1] 1.984467
```

**95% confidence interval** for  $\beta_2$  is  $0.567 \pm 1.98 \times 0.0793 = [0.410, 0.724]$

# Confidence Intervals

We have a confidence interval for  $\beta_2$ , i.e.,  $[0.410, 0.724]$ .

## What does it mean?

**Informally:** The confidence interval gives us a region (interval) in which we can place some trust (confidence) for containing the parameter.

**More formally:** If we repeatedly sample from our population and construct confidence intervals for each of these samples, then  $(1 - \alpha) \cdot 100$  percent of our intervals (e.g., 95%) will contain the population parameter *somewhere in the interval*.

Now back to our simulation...

# Confidence Intervals

We drew 10,000 samples (each of size  $n = 30$ ) from our population and estimated our regression model for each sample:

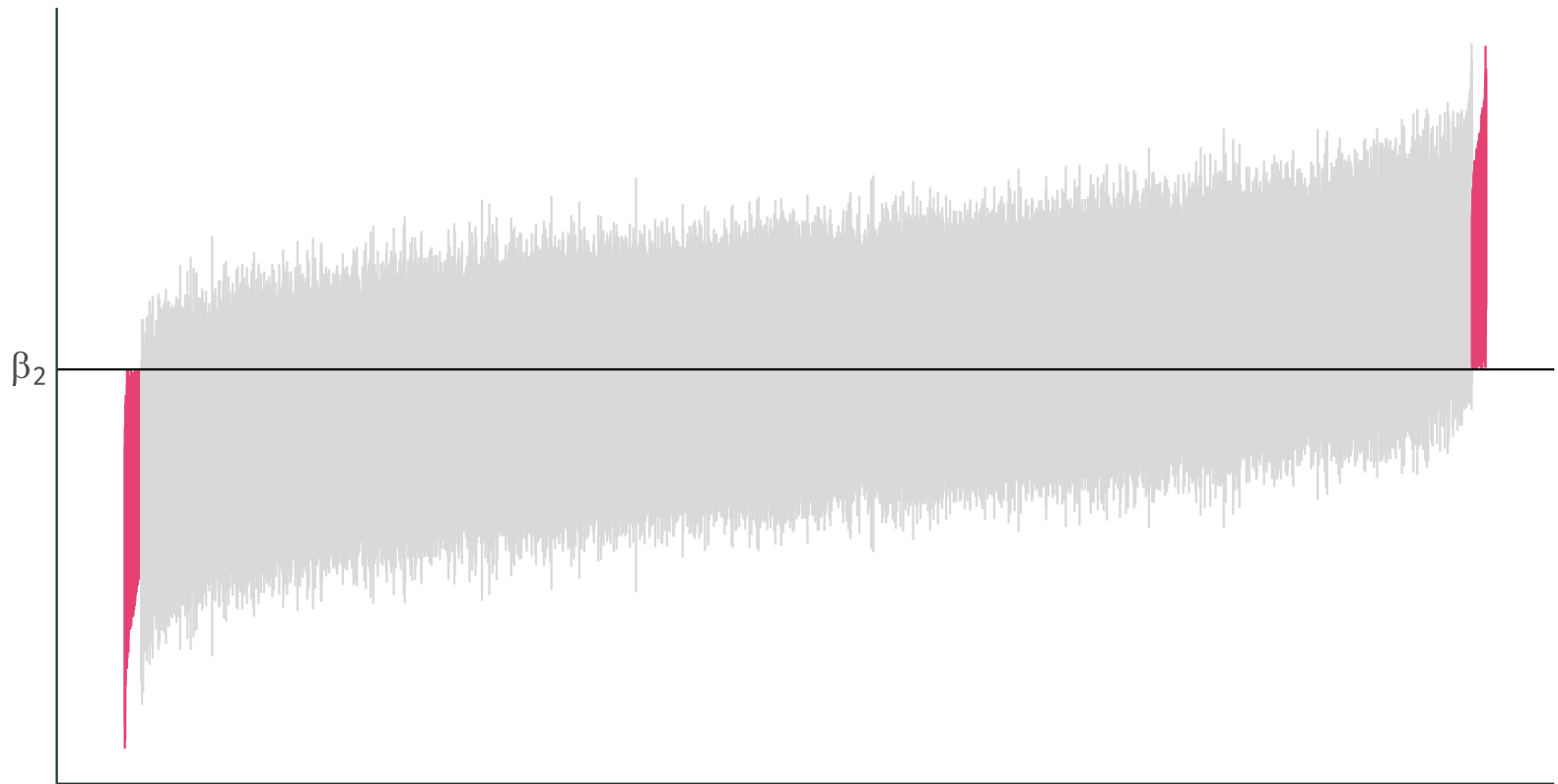
$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

(repeated 10,000 times)

Now, let's estimate 95% confidence intervals for each of these intervals...

# Confidence Intervals

**From our previous simulation:** 97.7% of 95% confidence intervals contain the true parameter value of  $\beta_2$ .



# Confidence Intervals

## Example: Association of police with crime

You can instruct `tidy` to return a 95 percent confidence interval for the association of campus police with campus crime:

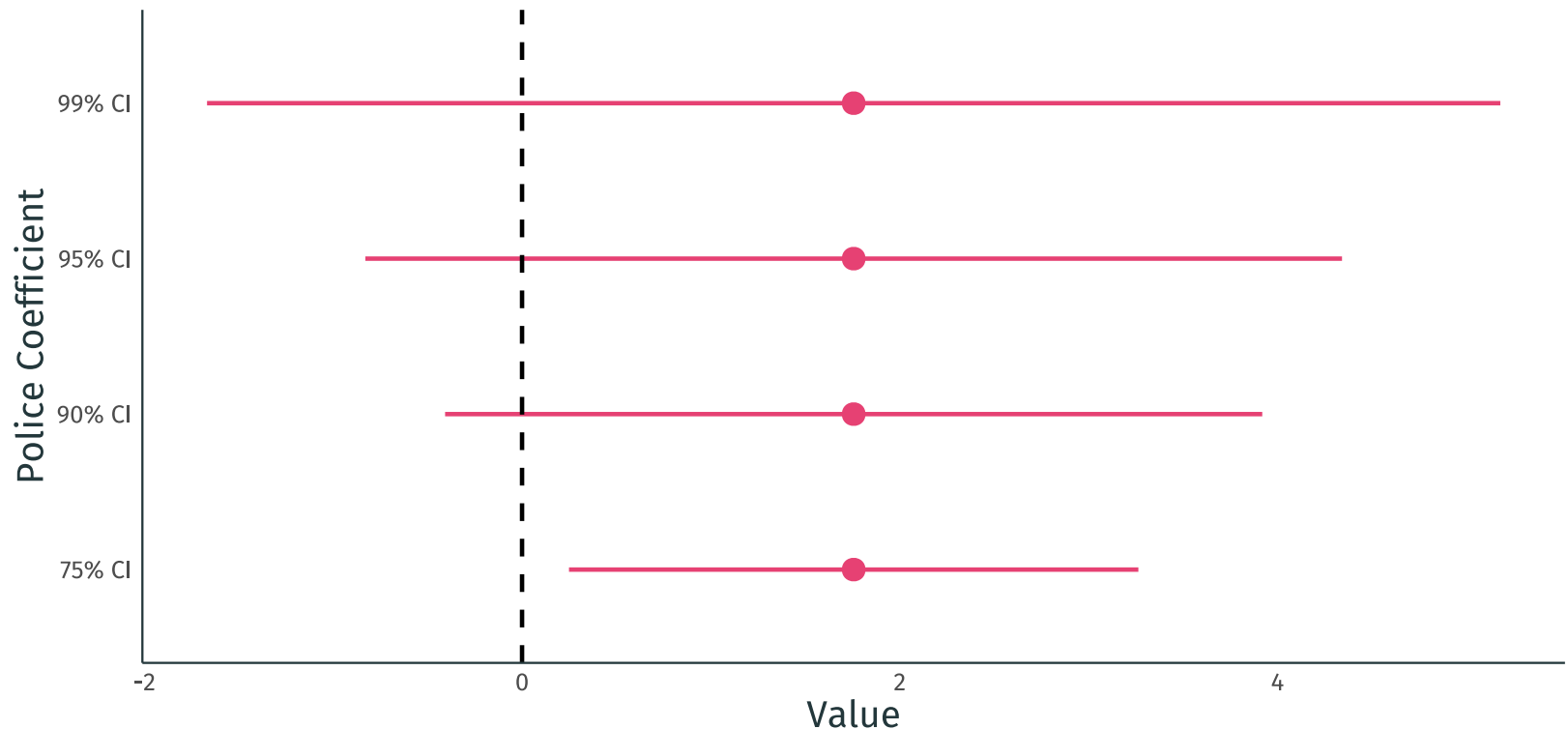
```
lm(crime ~ police, data = campus) %>% tidy(conf.int = TRUE, conf.level = 0.95)
```

```
#> # A tibble: 2 x 7
```

| #>   | term        | estimate | std.error | statistic | p.value  | conf.low | conf.high |
|------|-------------|----------|-----------|-----------|----------|----------|-----------|
| #>   | <chr>       | <dbl>    | <dbl>     | <dbl>     | <dbl>    | <dbl>    | <dbl>     |
| #> 1 | (Intercept) | 18.4     | 2.38      | 7.75      | 1.06e-11 | 13.7     | 23.1      |
| #> 2 | police      | 1.76     | 1.30      | 1.35      | 1.81e- 1 | -0.830   | 4.34      |

# Confidence Intervals

## Example: Association of police with crime



Four confidence intervals for the same coefficient.