2.6 — Statistical Inference

ECON 480 • Econometrics • Fall 2020

Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF20

metricsF20.classes.ryansafner.com

Outline

Why Uncertainty Matters

Confidence Intervals

Confidence Intervals Using the infer Package

Hypothesis Testing

Digression: p-Values and the Philosophy of Science

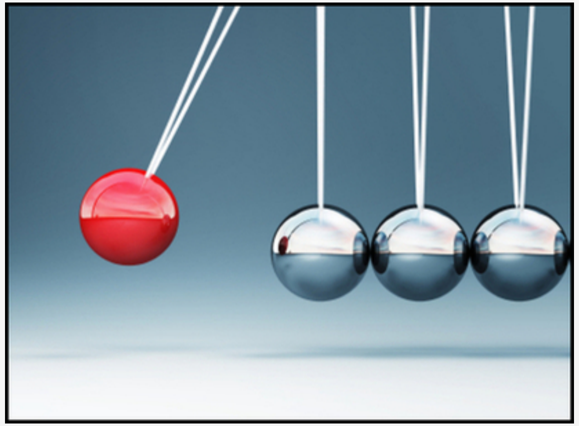
Why Uncertainty Matters

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

1. Problem for **identification**: **endogeneity**
   - X is **exogenous** if $cor(x, u) = 0$
   - X is **endogenous** if $cor(x, u) \neq 0$

2. Problem for **inference**: **randomness**
   - Data is random due to **natural sampling variation**
   - Taking one sample of a population will yield slightly different information than another sample of the same population



Distributions of the OLS Estimators

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a finite (specific) sample of data

- Our OLS model contains **2 sources of randomness**:

- ***Modeled* randomness**: u includes all factors affecting Y *other* than X
  - different samples will have different values of those other factors (u_i)

- ***Sampling* randomness**: different samples will generate different OLS estimators
  - Thus, $\hat{\beta}_0, \hat{\beta}_1$ are *also* **random variables**, with their own **sampling distribution**

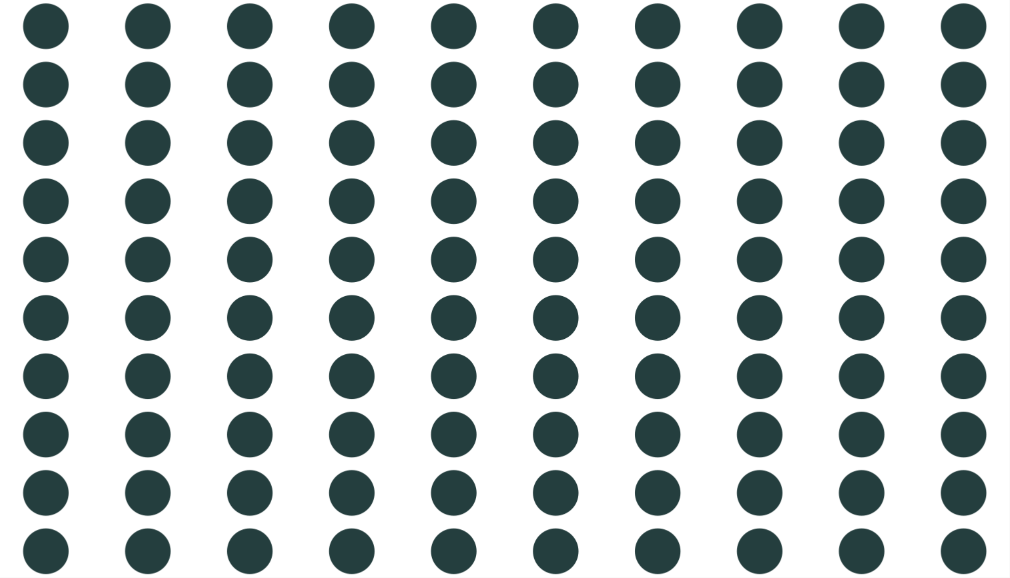
The Two Problems: Where We're Heading...Ultimately

$$\textbf{Sample} \underbrace{\rightarrow}_{\text{statistical inference}} \textbf{Population} \underbrace{\rightarrow}_{\text{causal indentification}} \textbf{Unobserved Parameters}$$

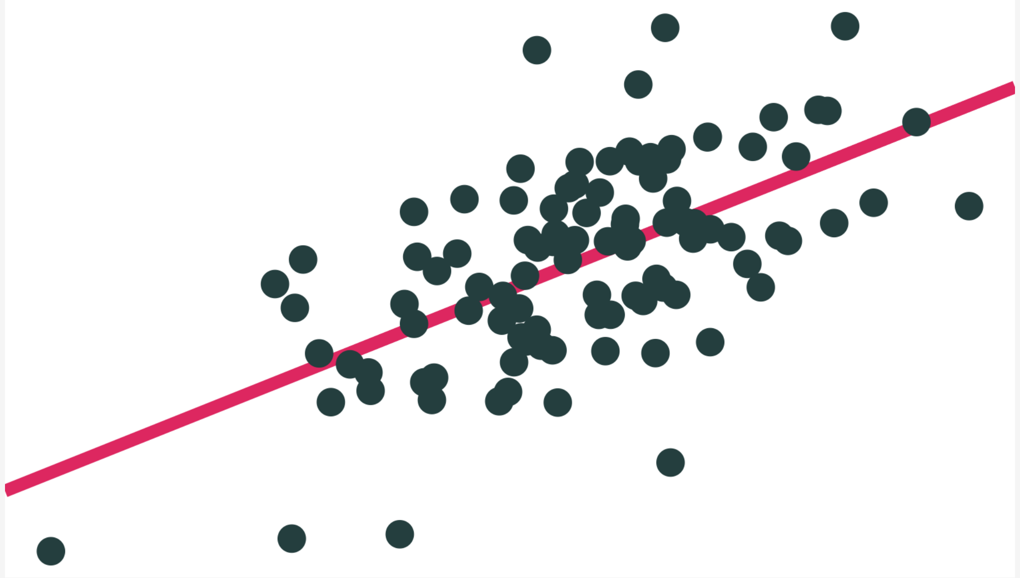
- We want to **identify** causal relationships between **population** variables
  - Logically first thing to consider
  - **Endogeneity problem**

- We'll use **sample** *statistics* to **infer** something about population *parameters*
  - In practice, we'll only ever have a finite *sample distribution* of data
  - We *don't* know the *population distribution* of data
  - **Randomness problem**

Why Sample vs. Population Matters



Population

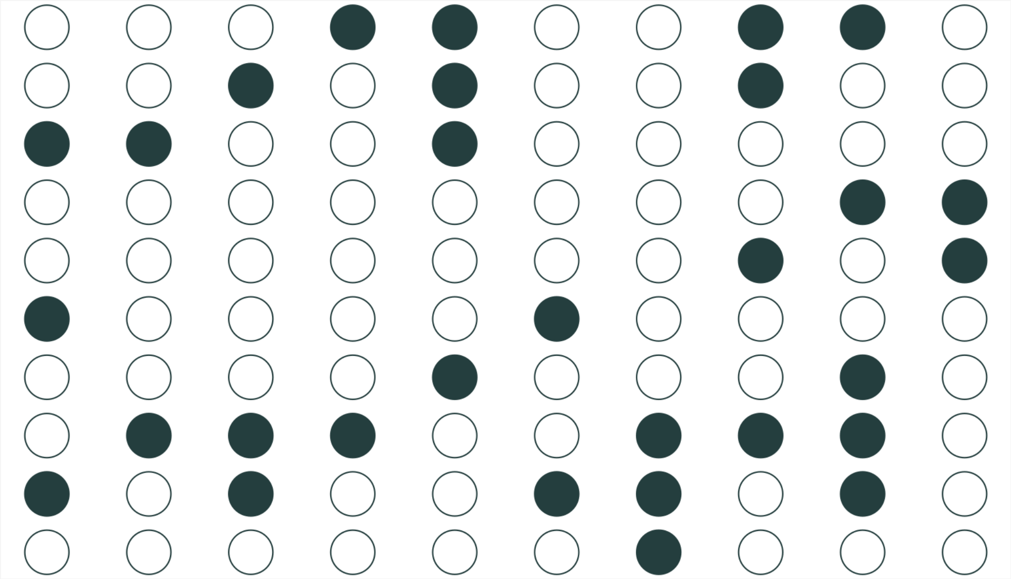


Population relationship

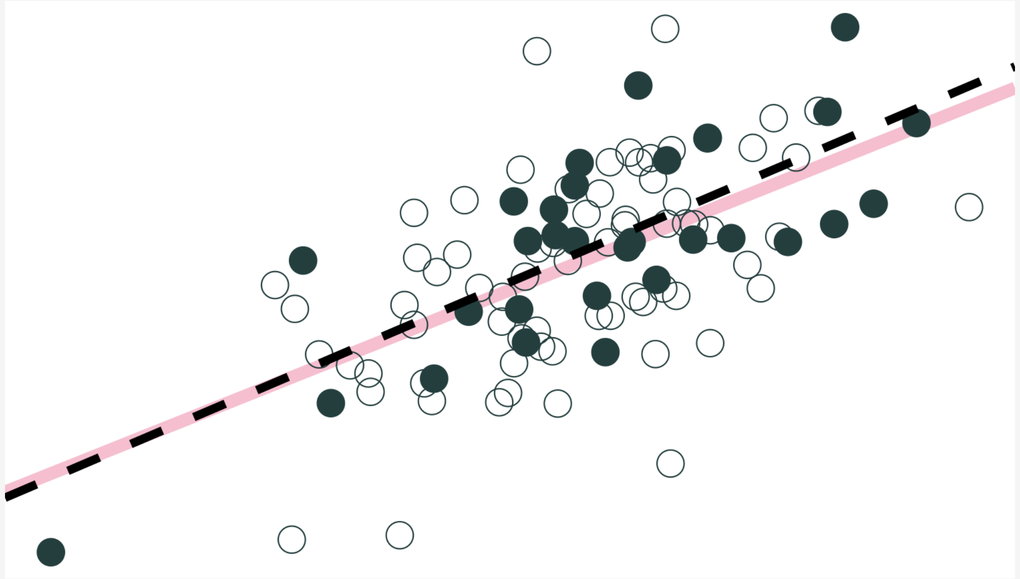
$$Y_i = 3.24 + 0.44X_i + u_i$$

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Why Sample vs. Population Matters



Sample 1: 30 random individuals



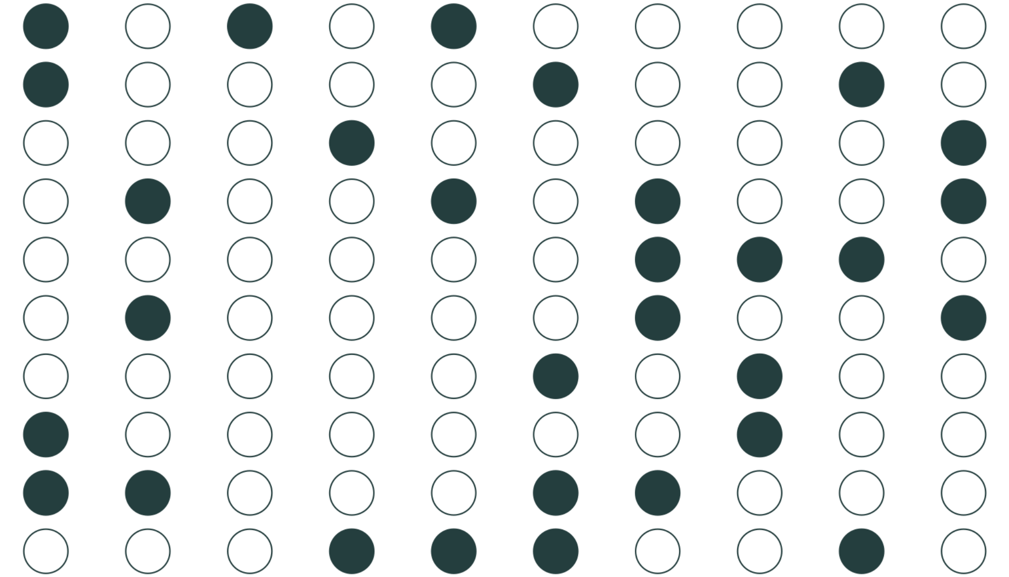
Population relationship

$$Y_i = 3.24 + 0.44X_i + u_i$$

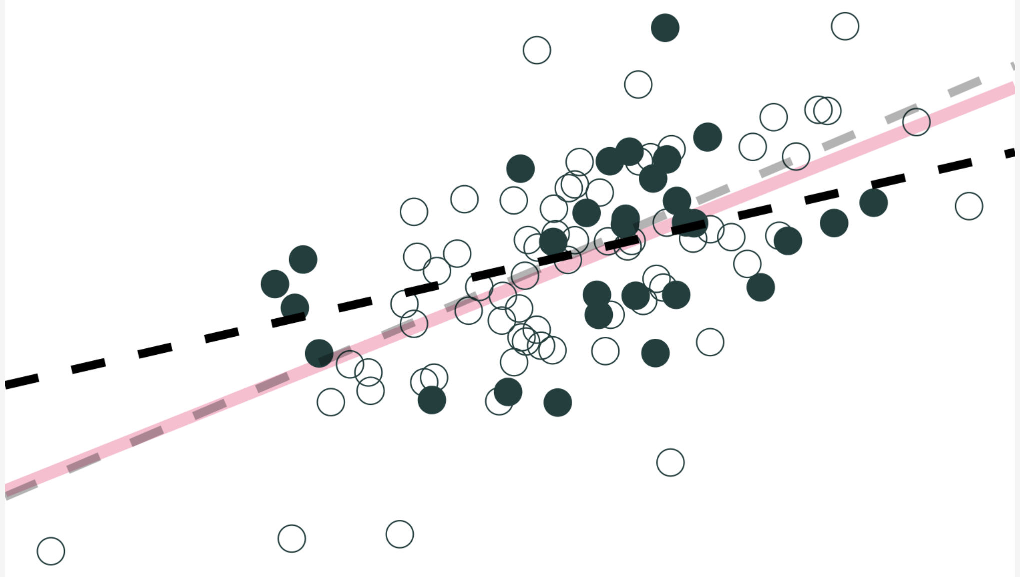
Sample relationship

$$\hat{Y}_i = 3.19 + 0.47X_i$$

Why Sample vs. Population Matters



Sample 2: 30 random individuals



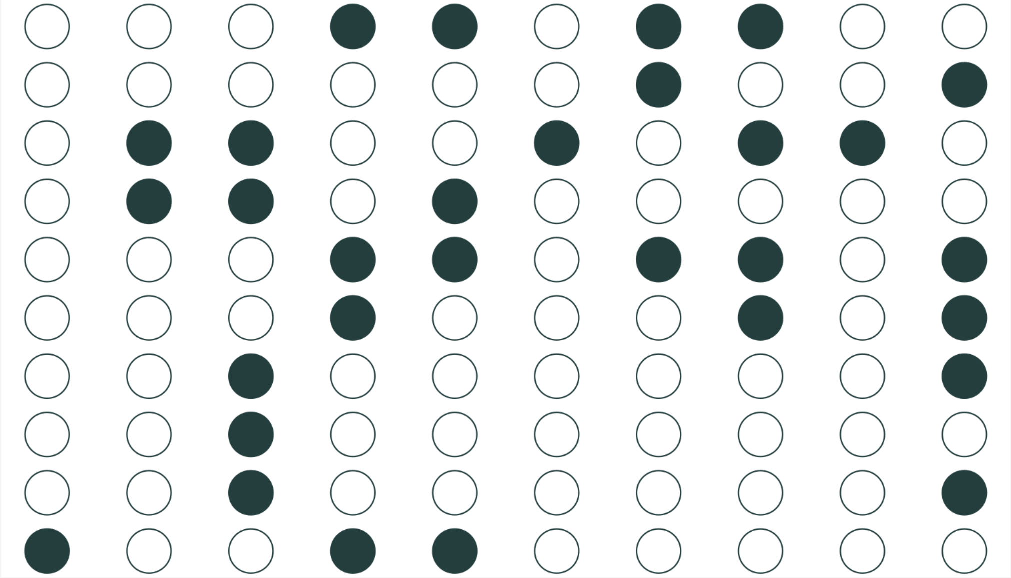
Population relationship

$$Y_i = 3.24 + 0.44X_i + u_i$$

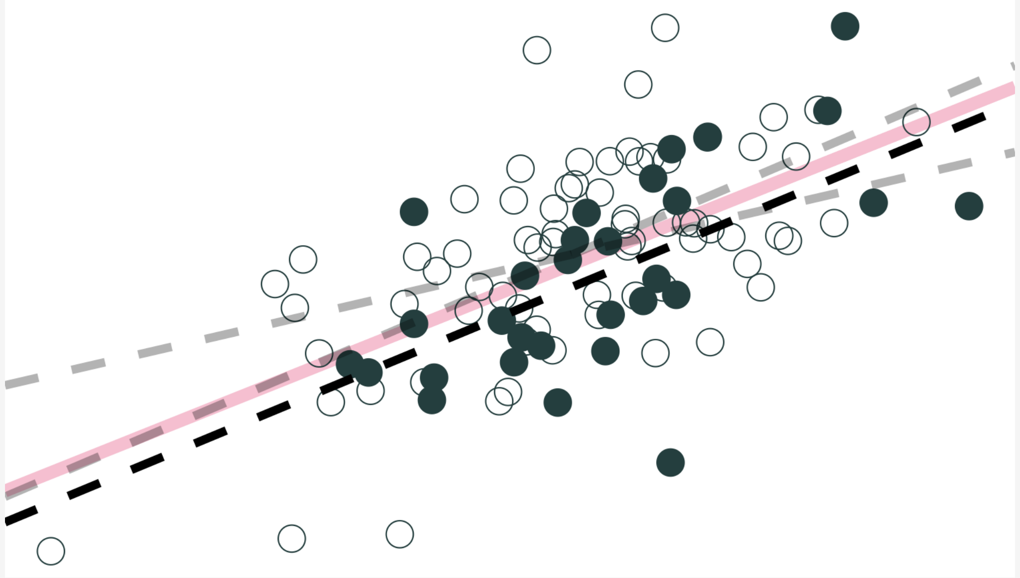
Sample relationship

$$\hat{Y}_i = 4.26 + 0.25X_i$$

Why Sample vs. Population Matters



Sample 3: 30 random individuals



Population relationship

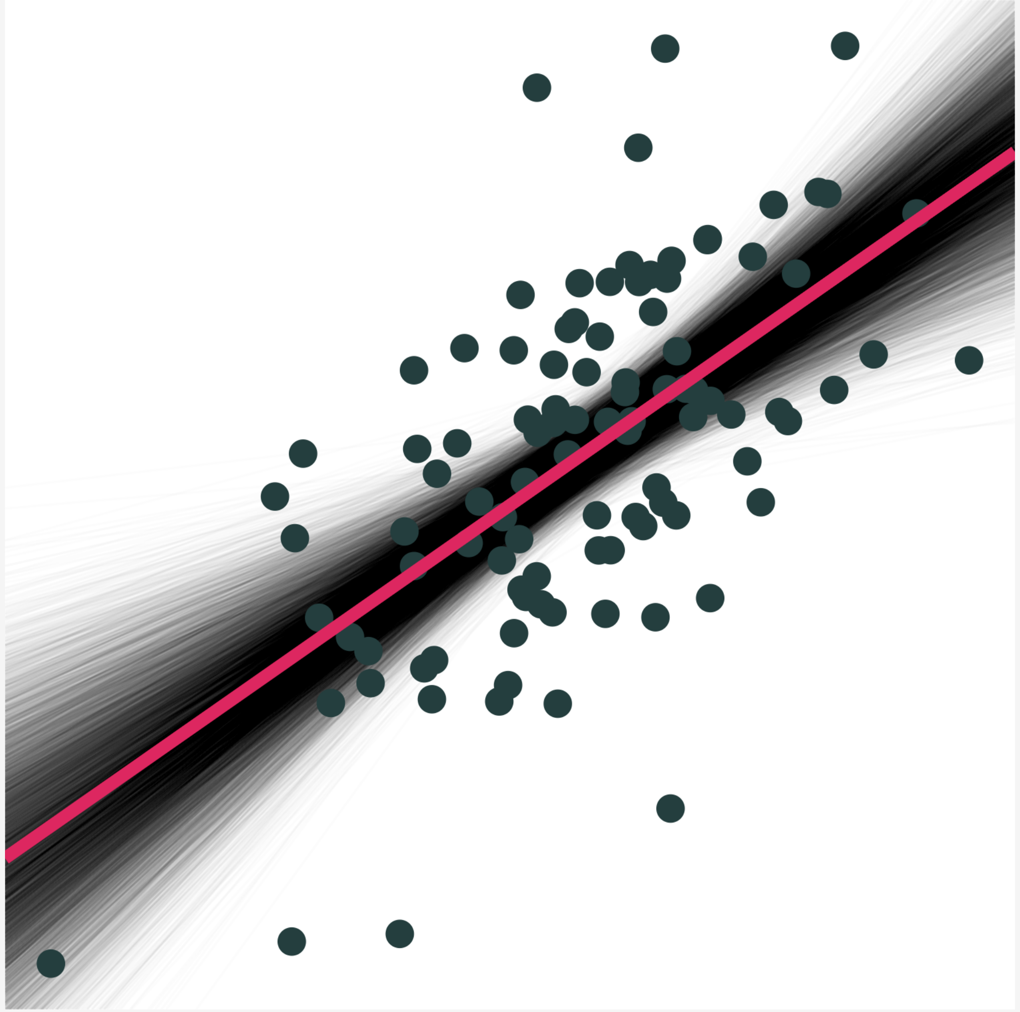
$$Y_i = 3.24 + 0.44X_i + u_i$$

Sample relationship

$$\hat{Y}_i = 2.91 + 0.46X_i$$

Why Sample vs. Population Matters

- Let's repeat this process **10,000 times**!
- This exercise is called a **(Monte Carlo) simulation**
  - I'll show you how to do this next class with the `infer` package



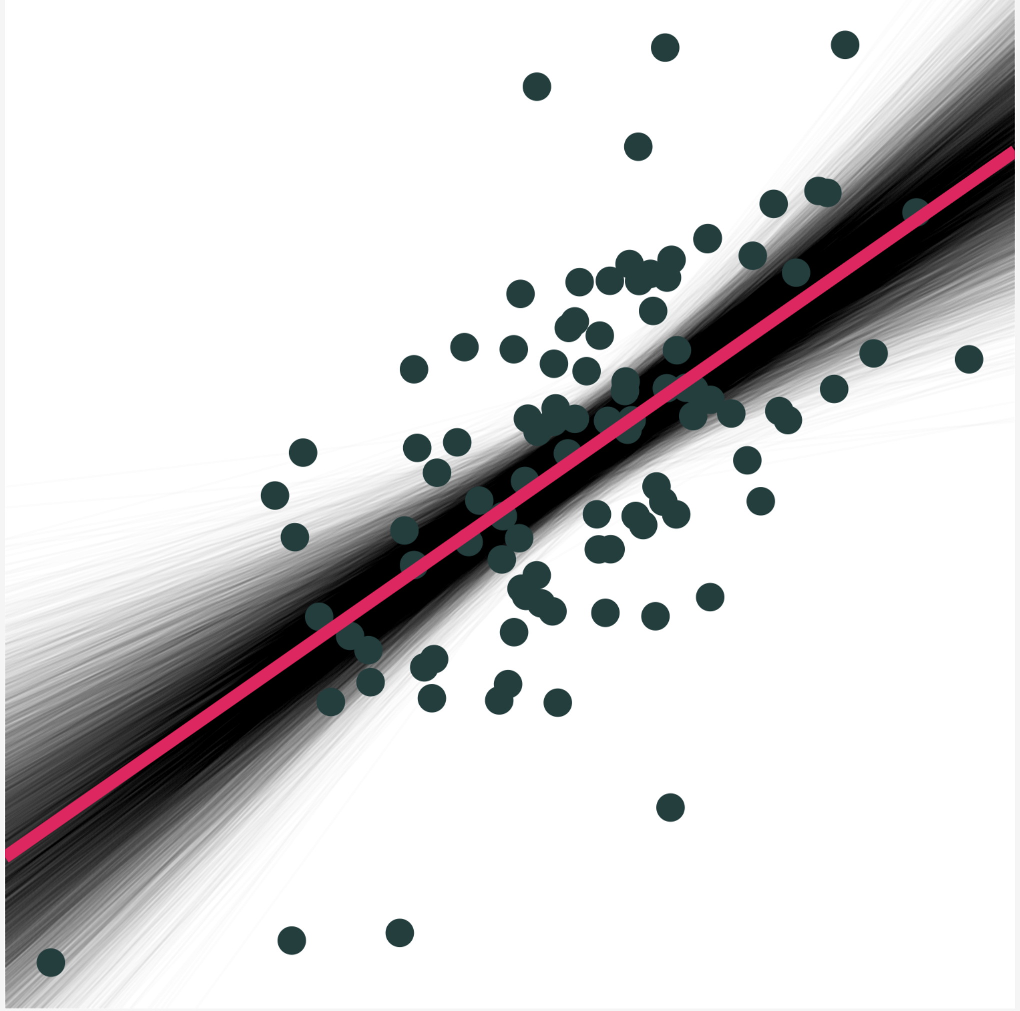
Why Sample vs. Population Matters

- On average estimated regression lines from our hypothetical samples provide an unbiased estimate of the true population regression line

$$E[\hat{\beta_1}] = \beta_1$$

- However, any *individual line* (any *one* sample) can miss the mark

- This leads to uncertainty about our estimated regression line
    - Remember, we only have *one* sample in reality!
    - This is why we care about the standard error of our line: $se(\hat{\beta_1})$!



Confidence Intervals

Statistical Inference

Sample $\xrightarrow{\text{statistical inference}}$ **Population** $\xrightarrow{\text{causal indentification}}$ **Unobserved Parameters**

Statistical Inference

$$\text{Sample} \xrightarrow{\text{statistical inference}} \text{Population} \xrightarrow{\text{causal indentification}} \textbf{Unobserved Parameters}$$

- So what we naturally want to start doing is **inferring** what the true population regression model is, using our estimated regression model from our sample

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X \xrightarrow{\text{🤞 hopefully 🤞}} Y_i = \beta_0 + \beta_1 X + u_i$$

- We can’t yet make causal inferences about whether/how X *causes* Y
  - coming after the midterm!

Estimation and Statistical Inference

- Our problem with **uncertainty** is we don't know whether our sample estimate is *close* or *far* from the unknown population parameter

- But we can use our errors to learn how well our model statistics likely estimate the true parameters

- Use $\hat{\beta_1}$ and its standard error, $se(\hat{\beta_1})$ for statistical inference about true β_1

- We have two options...

Estimation and Statistical Inference

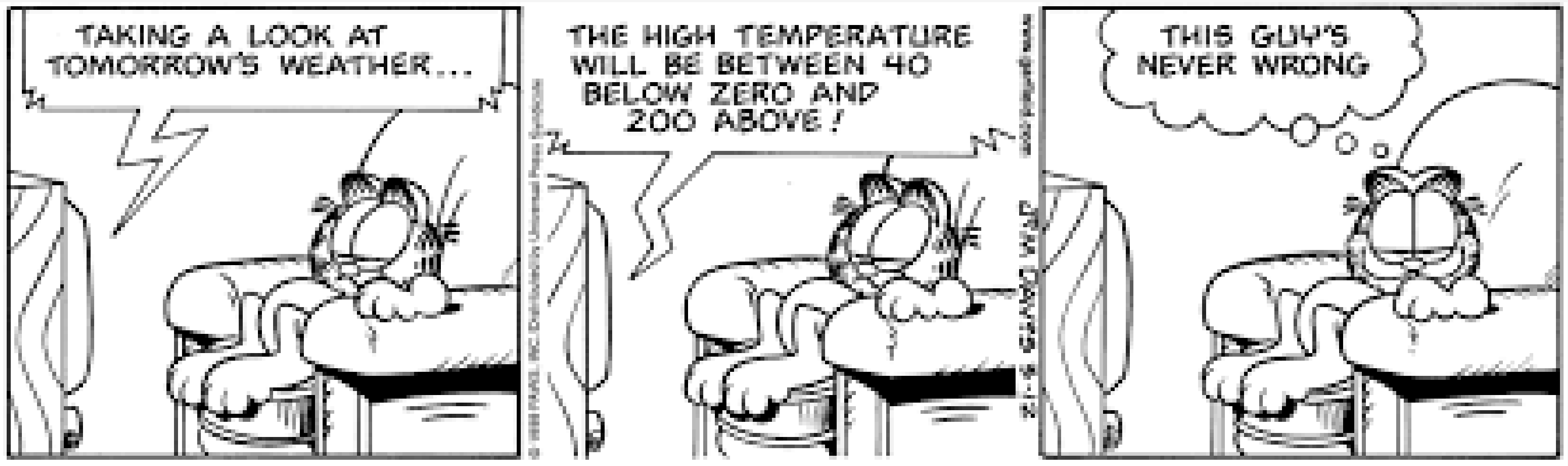
Point estimate

- Use our $\hat{\beta_1}$ and $se(\hat{\beta_1})$ to determine whether we have statistically significant evidence to reject a hypothesized β_1

Confidence interval

- Use $\hat{\beta_1}$ and $se(\hat{\beta_1})$ to create an *range* of values that gives us a good chance of capturing the true β_1

Accuracy vs. Precision


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG

- More typical in econometrics to do hypothesis testing (next class)

Generating Confidence Intervals

- We can generate our confidence interval by generating a **"bootstrap"** sampling distribution

- This takes our sample data, and resamples it by selecting random observations with replacement

- This allows us to approximate the sampling distribution of $\hat{\beta}_1$ by simulation!

Confidence Intervals Using the infer Package

Confidence Intervals Using the infer Package

- The `infer` package allows you to do statistical inference in a `tidy` way, following the philosophy of the `tidyverse`

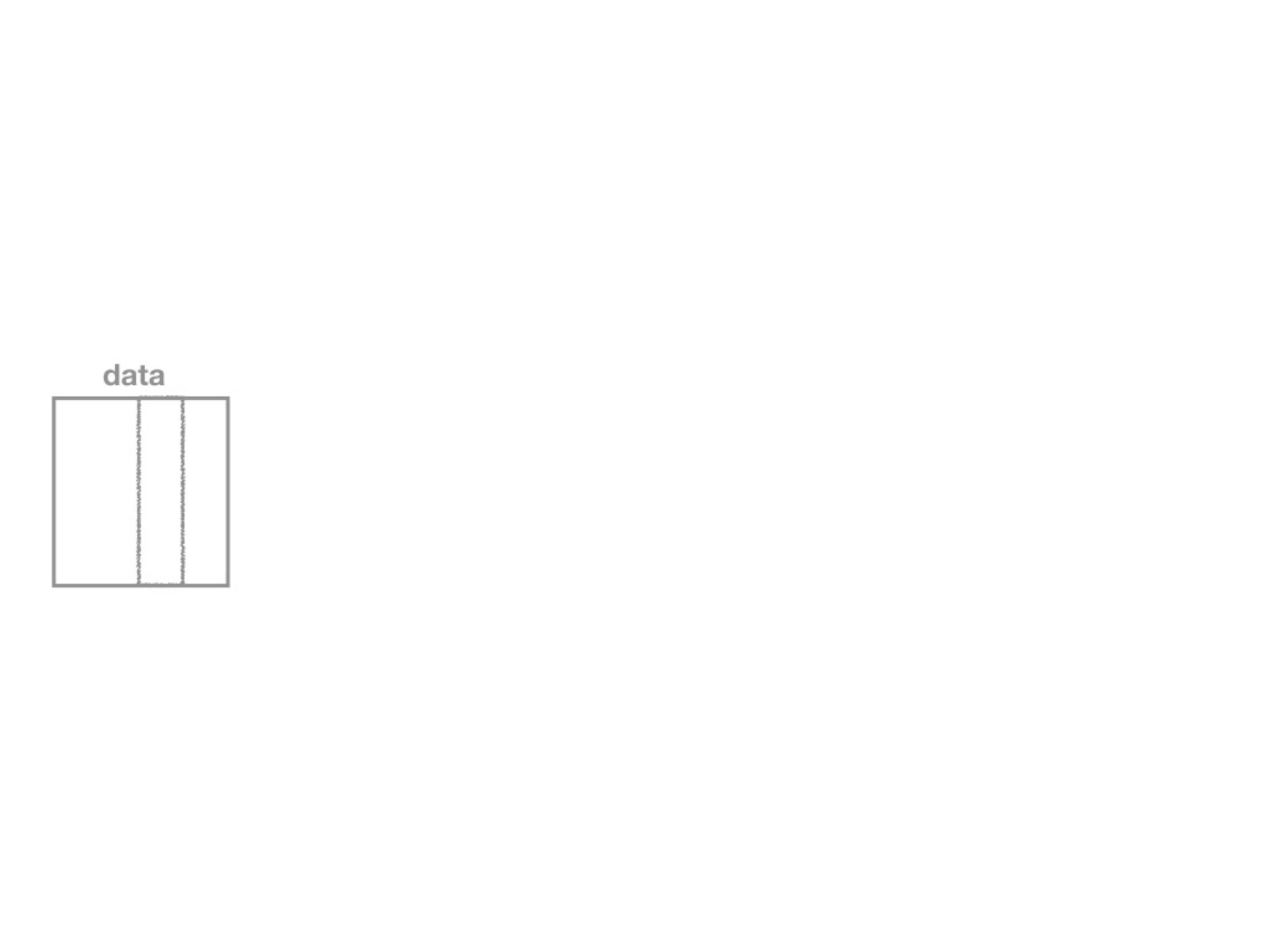
```r
# install first!
install.packages("infer")

# load
library(infer)
```

Confidence Intervals with the infer Package I

- `infer` allows you to run through these steps manually to understand the process:

1. `specify()` a model
2. `generate()` a bootstrap distribution
3. `calculate()` the confidence interval
4. `visualize()` with a histogram (optional)

Confidence Intervals with the infer Package II

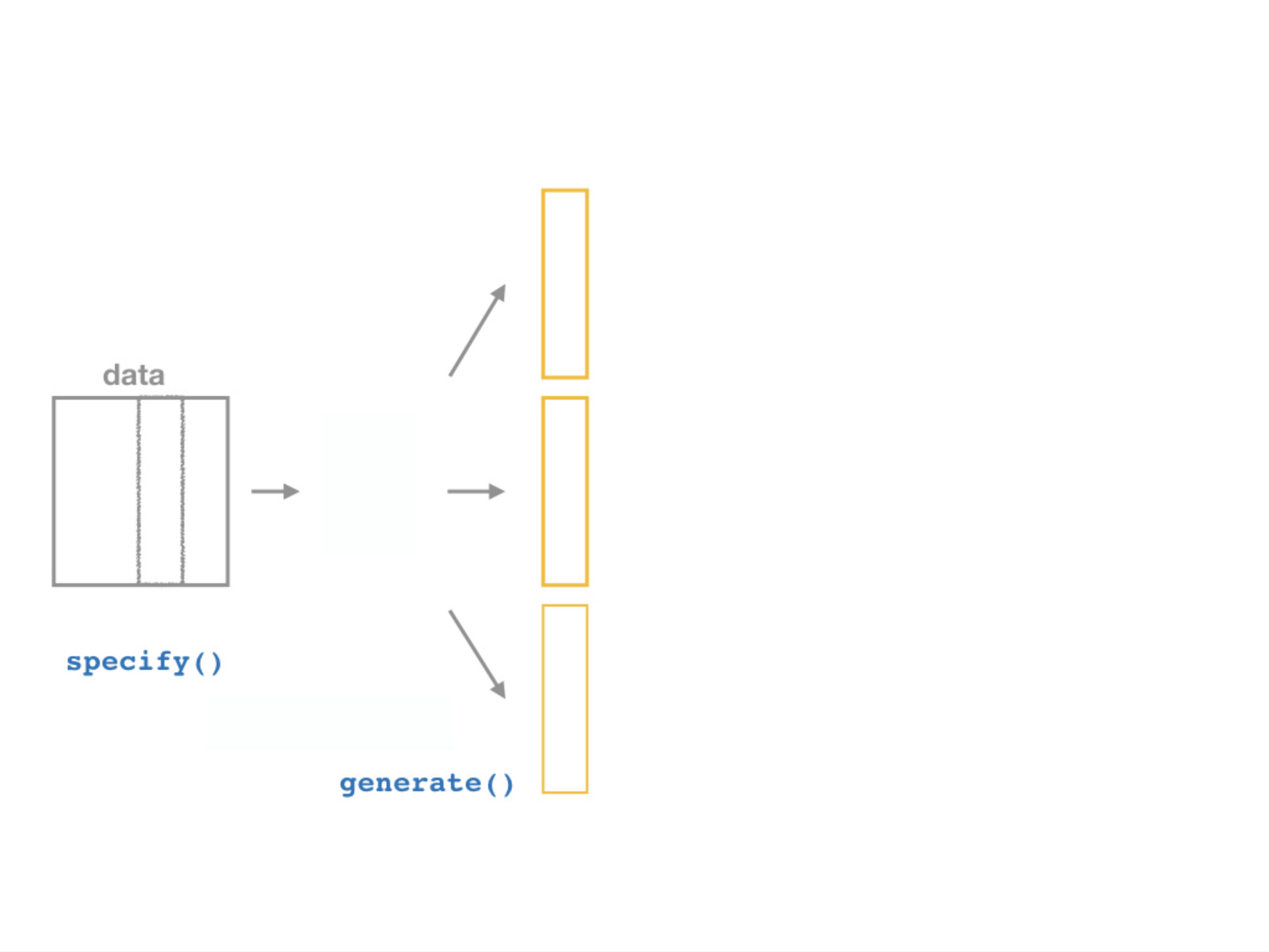


Confidence Intervals with the infer Package II

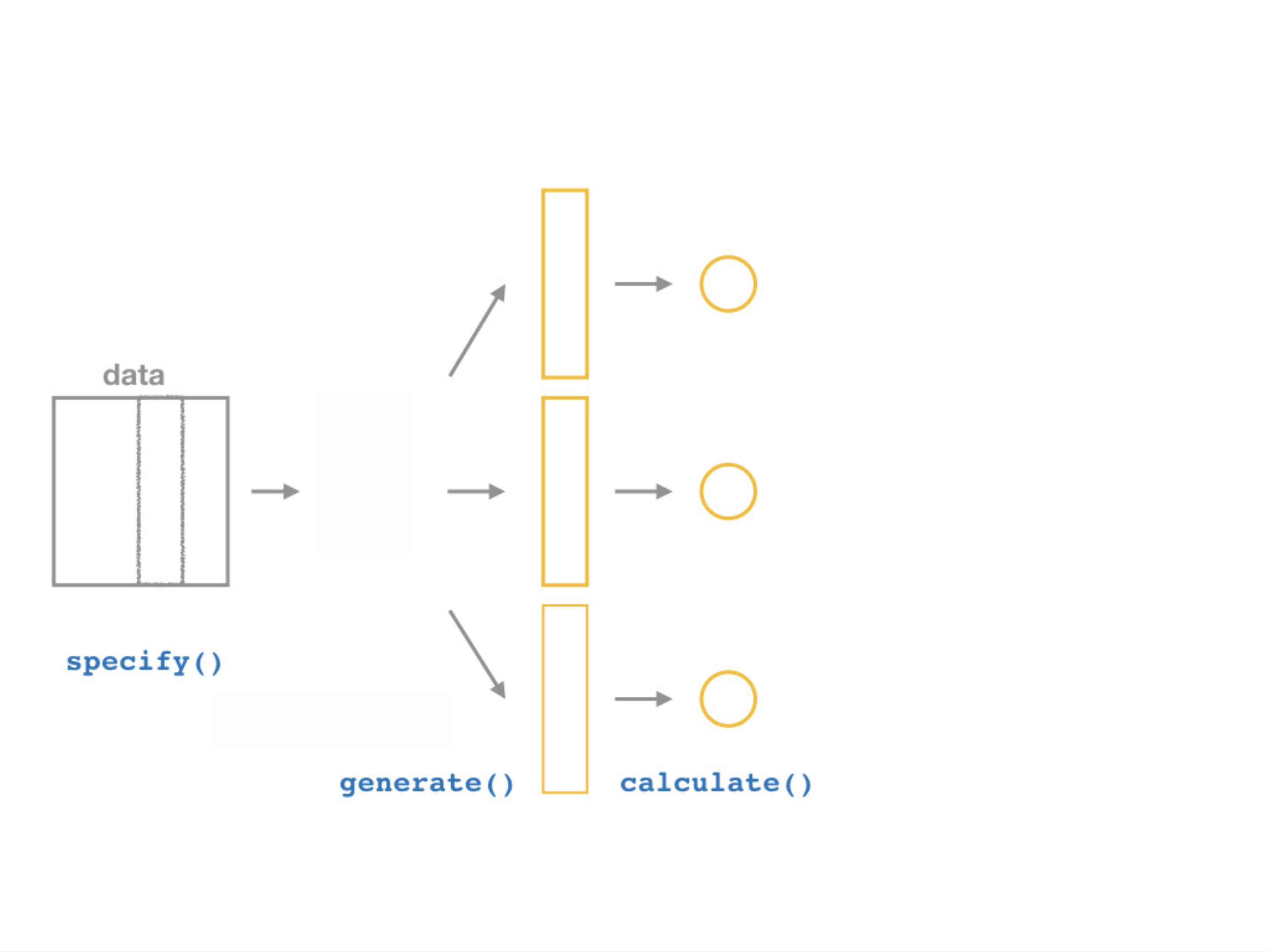
data

`specify()`

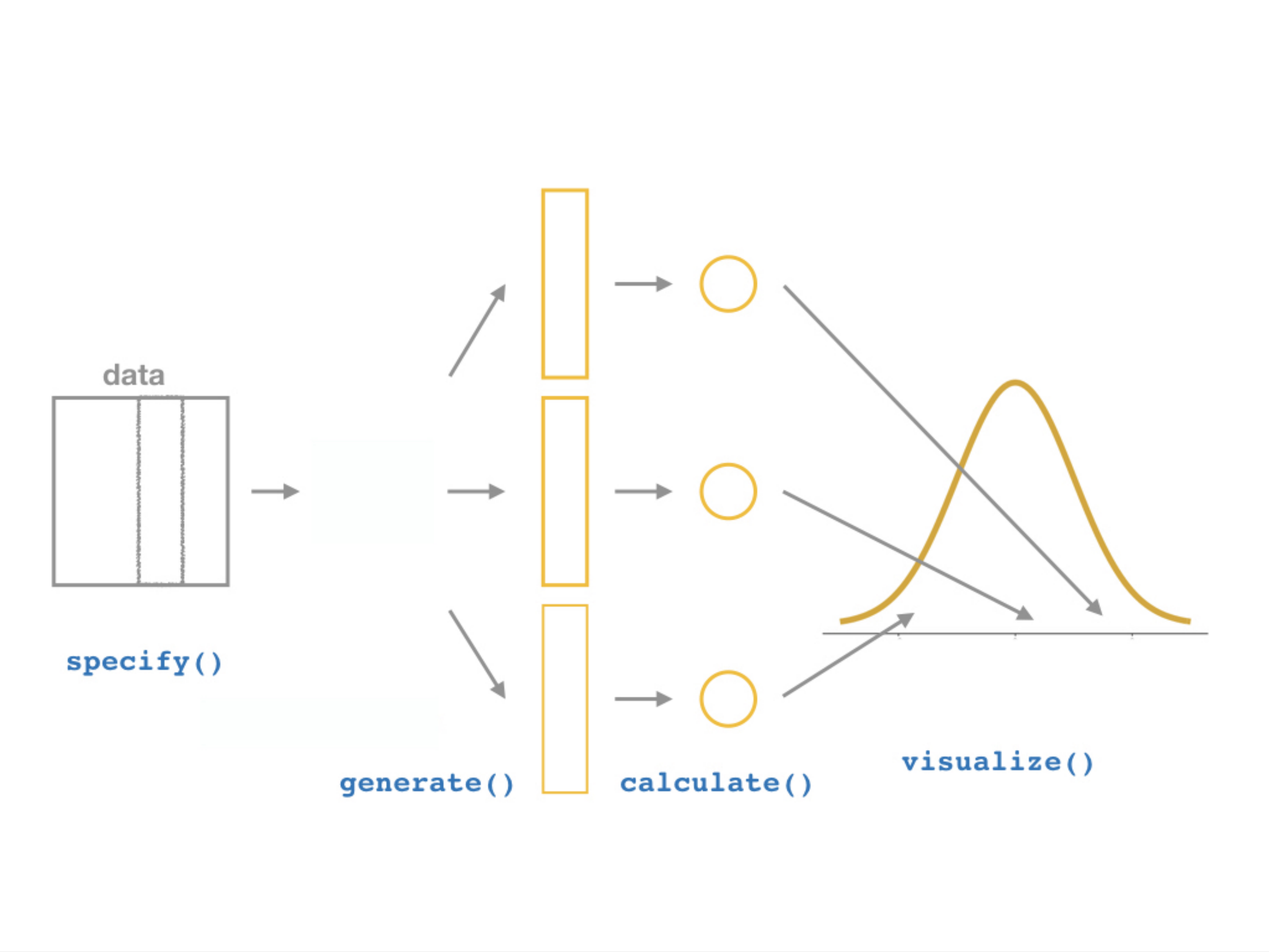
Confidence Intervals with the infer Package II



Confidence Intervals with the infer Package II



Confidence Intervals with the infer Package II



Bootstrapping

Our Sample

term <chr>	estimate <dbl>	std.error <dbl>
(Intercept)	698.932952	9.4674914
str	-2.279808	0.4798256

2 rows | 1-3 of 5 columns

Another “Sample”

term <chr>	estimate <dbl>	std.error <dbl>
(Intercept)	708.270835	9.5041448
str	-2.797334	0.4802065

2 rows | 1-3 of 5 columns

👆 Bootstrapped from Our Sample

- Now we want to do this 1,000 times to simulate the unknown sampling distribution of $\hat{\beta}_1$

The *infer* Pipeline: Specify

data
specify()

The *infer* Pipeline: Specify

Specify

```
data %>% specify(y ~ x)
```

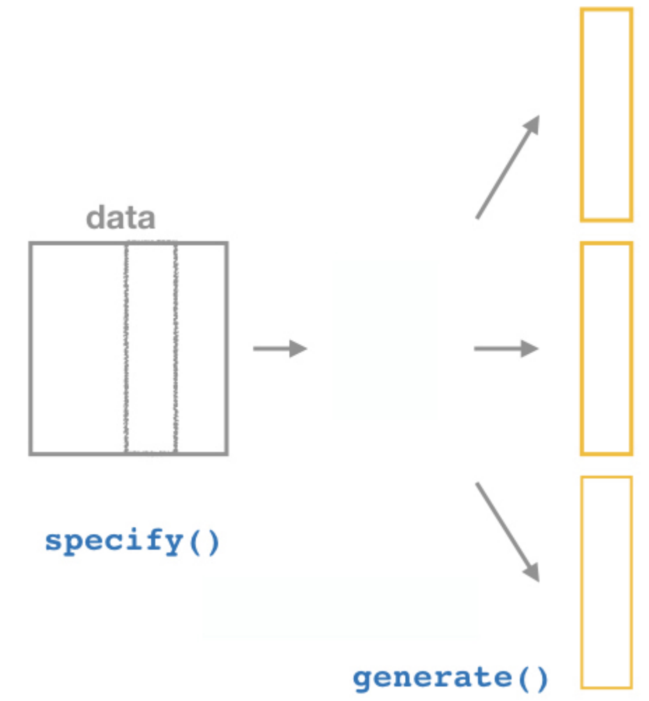
- Take our data and pipe it into the `specify()` function, which is essentially a `lm()` function for regression (for our purposes)

```
CASchool %>%
  specify(testscr ~ str)
```

testscr <dbl>	str <dbl>
690.80	17.88991
661.20	21.52466
643.60	18.69723
647.70	17.35714
640.85	18.67133

5 rows

The *infer* Pipeline: Generate



The *infer* Pipeline: Generate

Specify

Generate

`%>% generate(reps = n, type = "bootstrap")`

- Now the magic starts, as we run a number of simulated samples
- Set the number of `reps` and set `type` to `"bootstrap"`

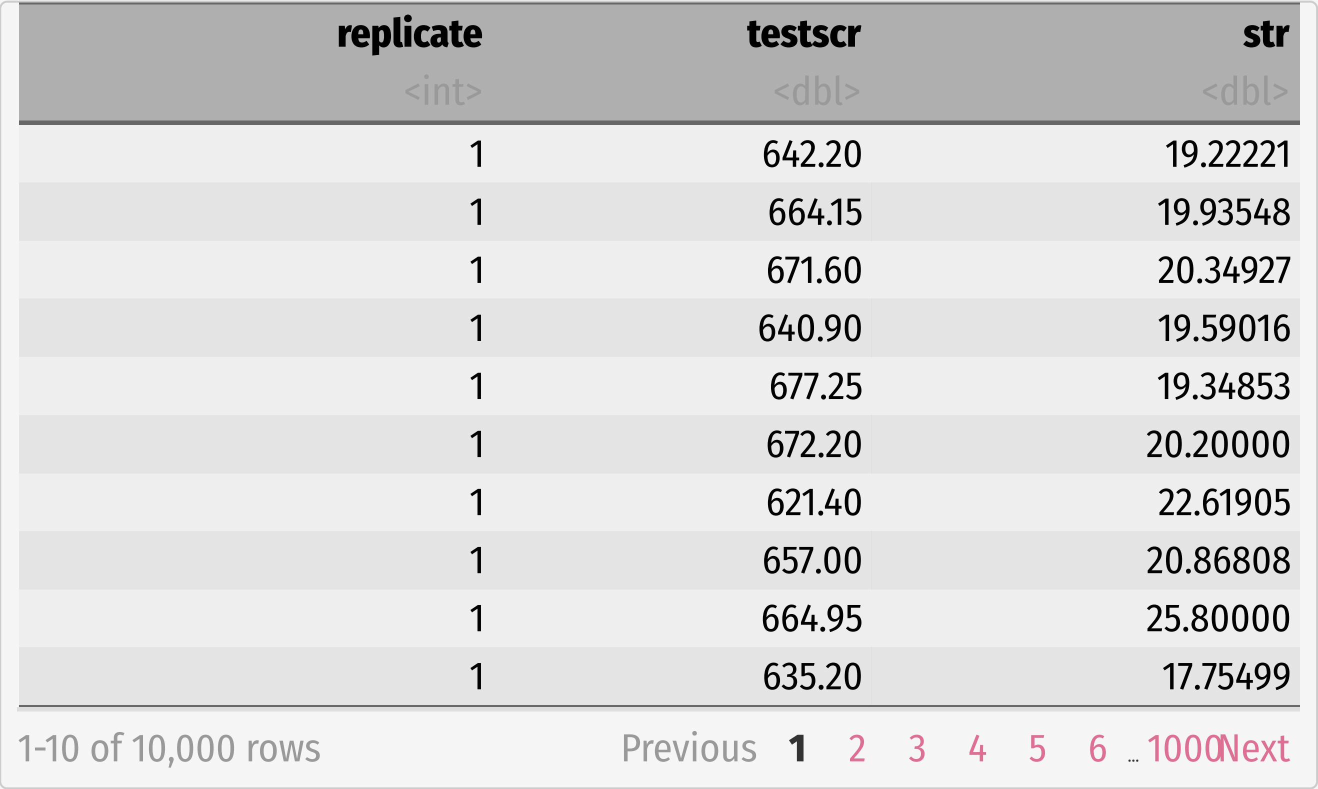
```
CASchool %>%
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap")
```

The *infer* Pipeline: Generate

Specify

Generate

```
%>% generate(reps = n,
type = "bootstrap")
```



replicate <int>	testscr <dbl>	str <dbl>
1	642.20	19.22221
1	664.15	19.93548
1	671.60	20.34927
1	640.90	19.59016
1	677.25	19.34853
1	672.20	20.20000
1	621.40	22.61905
1	657.00	20.86808
1	664.95	25.80000
1	635.20	17.75499

1-10 of 10,000 rows

- `replicate` : the “sample” number (1-1000)
- creates `x` and `y` values (data points)

The *infer* Pipeline: Calculate

Specify

Generate

Calculate

`%>% calculate(stat = "slope")`

```
CASchool %>%
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap") %>%
  calculate(stat = "slope")
```

- For each of the 1,000 replicates, calculate `slope` in `lm(testscr ~ str)`

- Calls it the `stat`

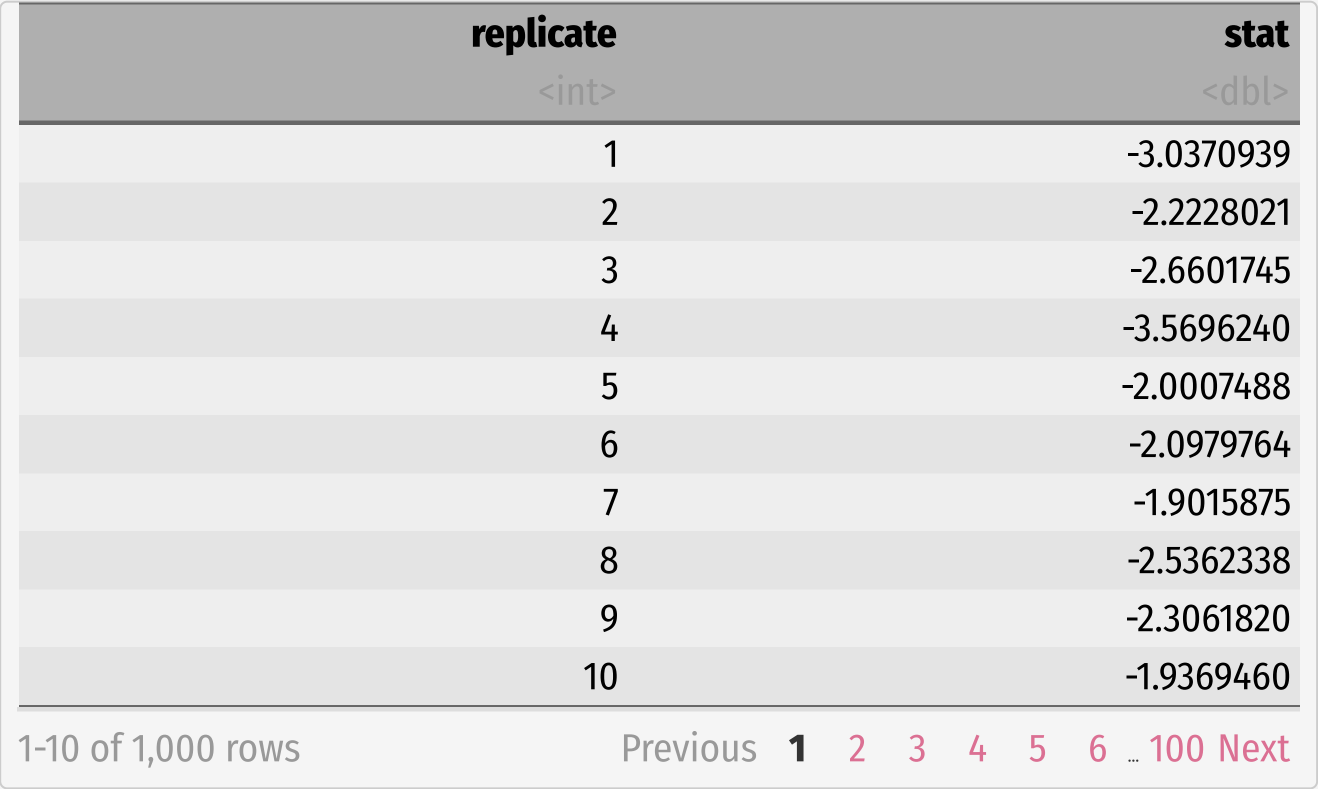
The *infer* Pipeline: Calculate

Specify

Generate

Calculate

```
%>% calculate(stat =
"slope")
```



replicate <int>	stat <dbl>
1	-3.0370939
2	-2.2228021
3	-2.6601745
4	-3.5696240
5	-2.0007488
6	-2.0979764
7	-1.9015875
8	-2.5362338
9	-2.3061820
10	-1.9369460

1-10 of 1,000 rows

Previous 1 2 3 4 5 6 ... 100 Next

The *infer* Pipeline: Calculate

Specify

Generate

Calculate

`%>% calculate(stat = "slope")`

```r
boot <- CASchool %>% #<< # save this
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap") %>%
  calculate(stat = "slope")
```

- `boot` is (our simulated) sampling distribution of $\hat{\beta}_1$!
- We can now use this to estimate the confidence interval from *our* $\hat{\beta}_1 = -2.28$
- And visualize it

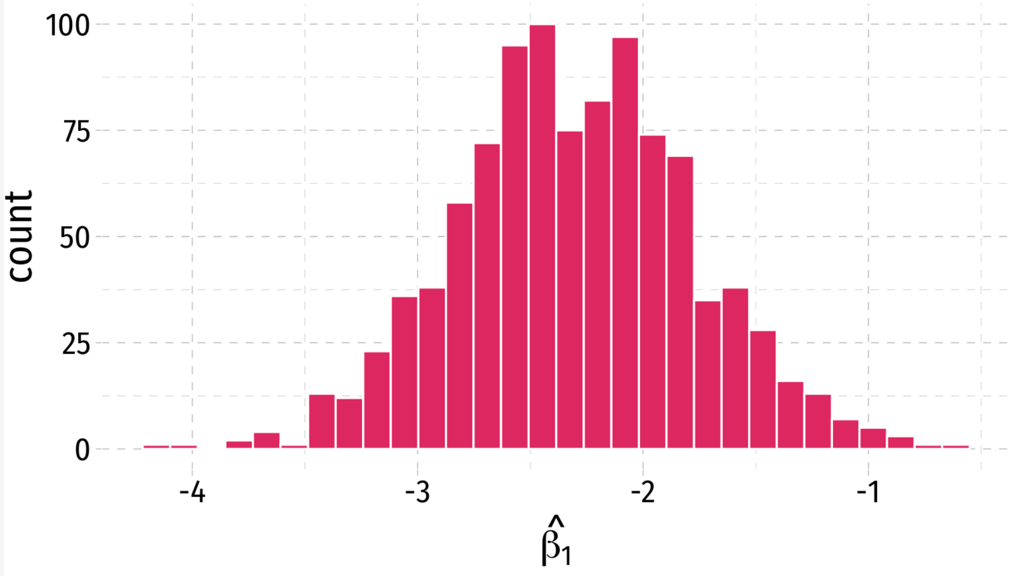
Confidence Interval

- A 95% confidence interval is the middle 95% of the sampling distribution

lower	upper
<dbl>	<dbl>
-3.340545	-1.238815

1 row

```
sampling_dist<-ggplot(data = boot)+
  aes(x = stat)+
  geom_histogram(color="white", fill = "#e64173
  labs(x = expression(hat(beta[1])))+
  theme_pander(base_family = "Fira Sans Condens
               base_size=20)
sampling_dist
```



Confidence Interval

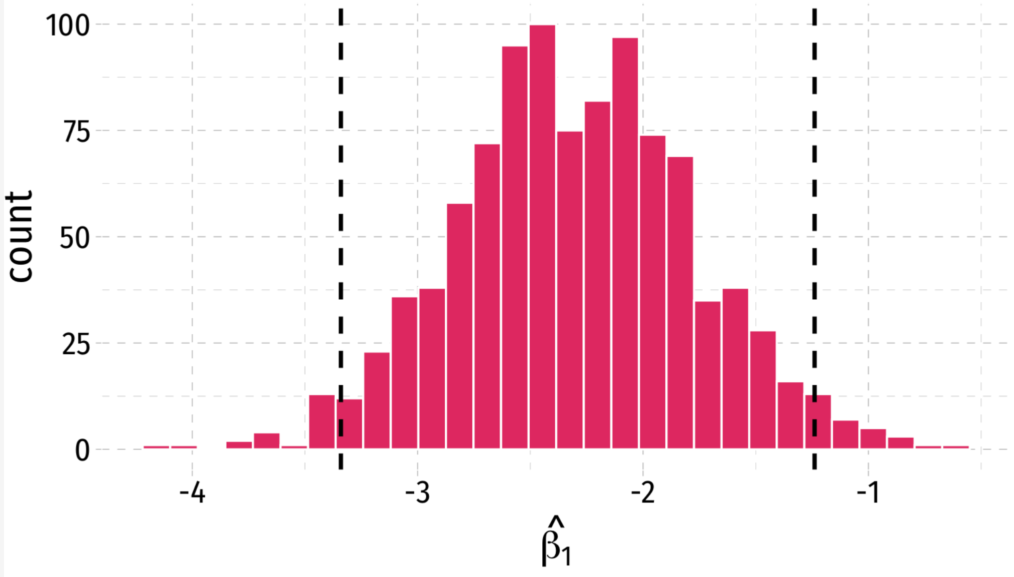
- A confidence interval is the middle 95% of the sampling distribution

```
ci<-boot %>%
  summarize(lower = quantile(stat, 0.025),
            upper = quantile(stat, 0.975))
ci
```

lower	upper
<dbl>	<dbl>
-3.340545	-1.238815

1 row

```
sampling_dist+
  geom_vline(data = ci, aes(xintercept = lower)
  geom_vline(data = ci, aes(xintercept = upper)
```



The *infer* Pipeline: Confidence Interval

Specify

Generate

Calculate

Get Confidence Interval

`%>%`
`get_confidence_interval()`

```
CASchool %>% #<< # save this
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap") %>%
  calculate(stat = "slope") %>%
  get_confidence_interval(level = 0.95,
                          type = "se",
                          point_estimate = -2.28)
```

lower_ci <dbl>	upper_ci <dbl>
-3.273376	-1.286624

1 row

Broom Can Estimate a Confidence Interval

```r
tidy_reg <- school_reg %>% tidy(conf.int = T)
tidy_reg
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	698.932952	9.4674914	73.824514	6.569925e-242	680.32313	717.542779
str	-2.279808	0.4798256	-4.751327	2.783307e-06	-3.22298	-1.336637

2 rows

```r
# save and extract confidence interval
our_CI <- tidy_reg %>%
  filter(term == "str") %>%
  select(conf.low, conf.high)

our_CI
```

conf.low	conf.high
<dbl>	<dbl>
-3.22298	-1.336637

1 row

The *infer* Pipeline: Confidence Interval

Specify

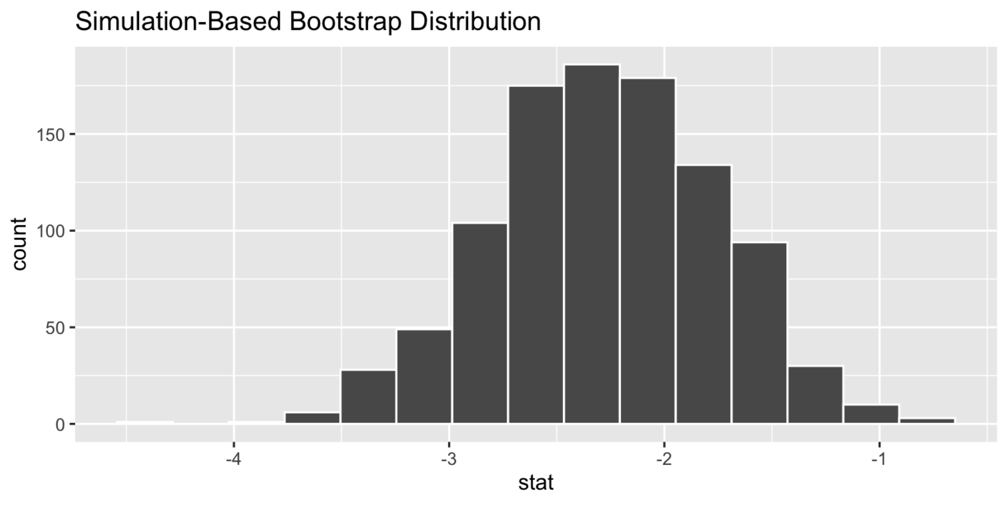
Generate

Calculate

Visualize

`%>% visualize()`

```
CASchool %>% #<< # save this
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap") %>%
  calculate(stat = "slope") %>%
  visualize()
```



- `visualize()` is just a wrapper for `ggplot()`

The *infer* Pipeline: Confidence Interval

Specify

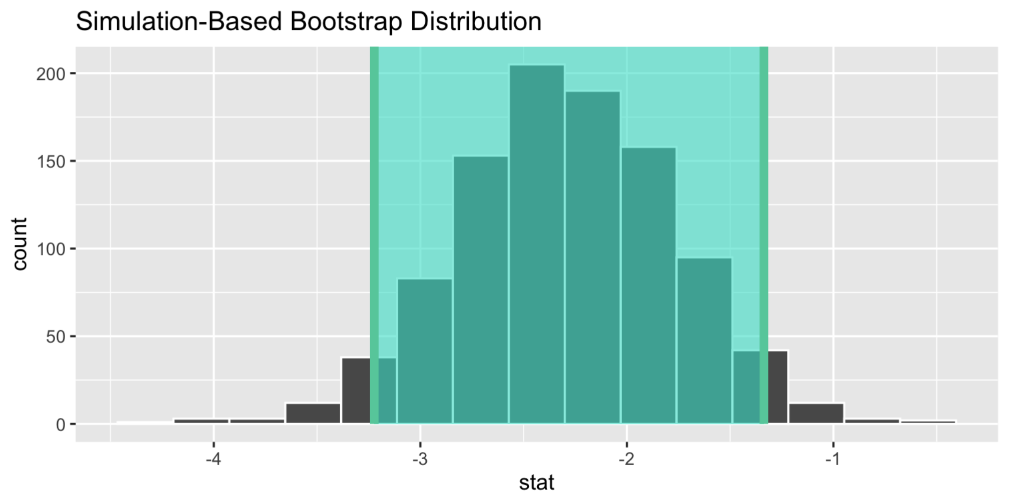
Generate

Calculate

Visualize

`%>% visualize()`

```
CASchool %>% #<< # save this
  specify(testscr ~ str) %>%
  generate(reps = 1000,
           type = "bootstrap") %>%
  calculate(stat = "slope") %>%
  visualize()+shade_ci(endpoints = our_CI)
```



Confidence Intervals

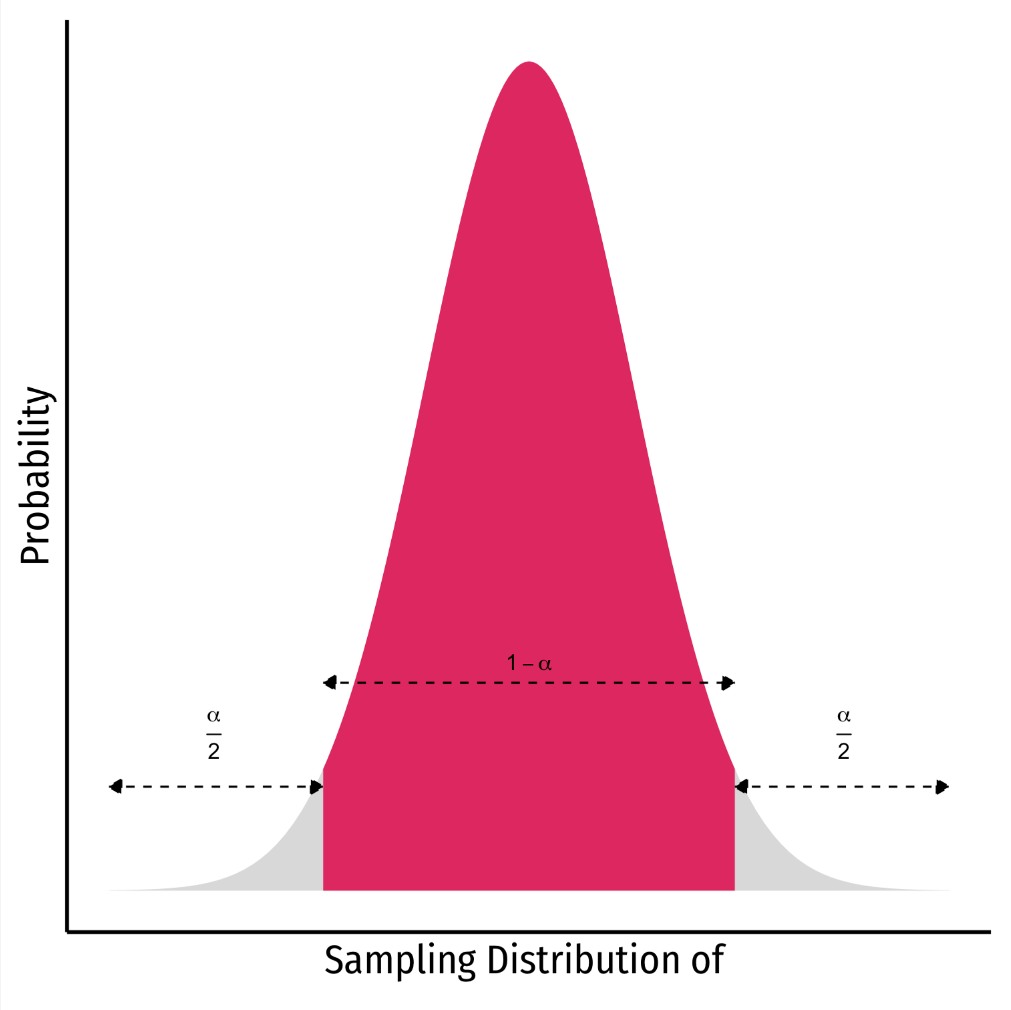
- In general, a **confidence interval (CI)** takes a point estimate and extrapolates it within some **margin of error**:

$$\left(\text{point estimate} - \text{margin of error}, \text{point estimate} + \text{margin of error} \right)$$

- The main question is, **how confident do we want to be** that our interval contains the true parameter?
  - Larger confidence level, larger margin of error (and thus larger interval)
- **$1- \alpha$** is the **confidence level** of our confidence interval
  - **α** is the **"significance level"** that we use in hypothesis testing
  - α: probability that the true mean is *not* contained within our interval
- Typical levels: 90%, 95%, 99%
  - 95% is especially common, $\alpha = 0.05$

Confidence Levels

- Depending on our confidence level, we are essentially looking for the center $(1-\alpha)$% of the sampling distribution
- Puts $\frac{\alpha}{2}$ in each tail

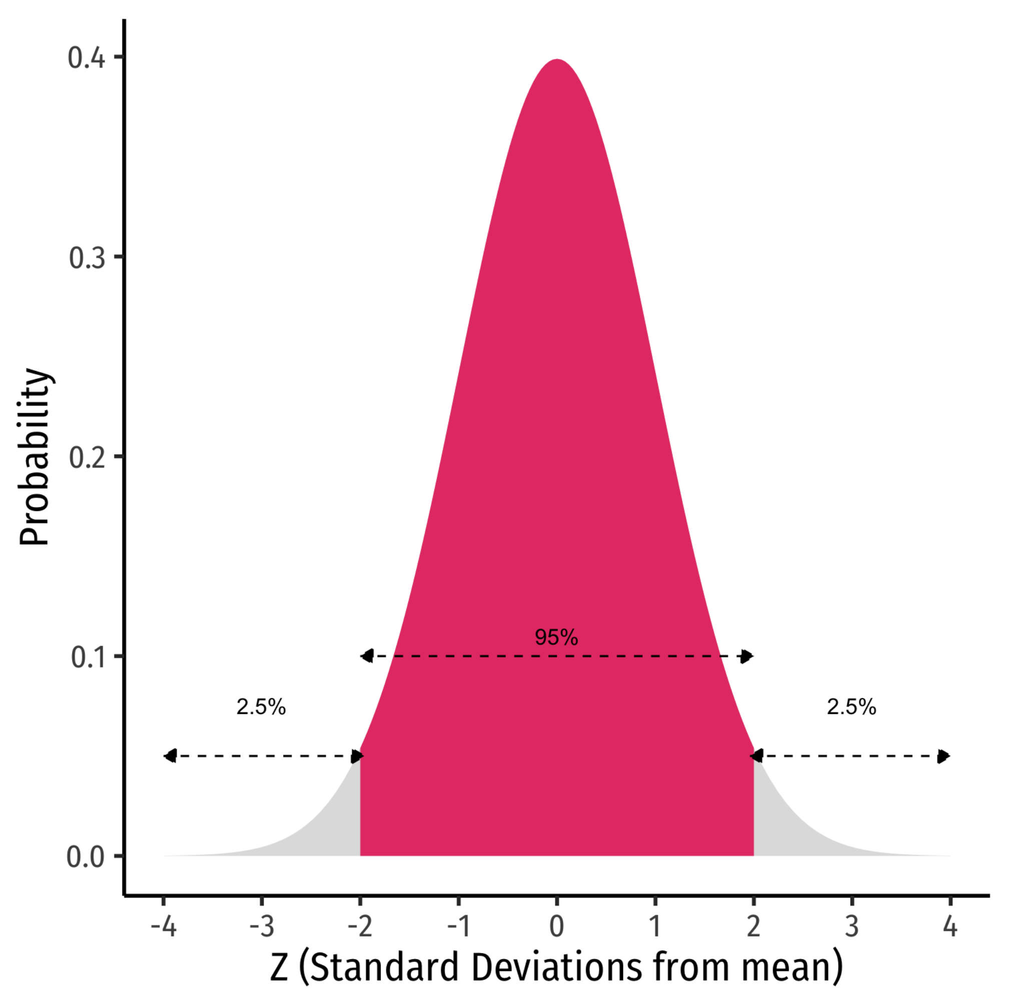


Confidence Levels and the Empirical Rule

- Recall the **68-95-99.7% empirical rule** for (standard) normal distributions![†]

- 95% of data falls within 2 standard deviations of the mean

- Thus, in 95% of samples, the true parameter is likely to fall within *about* 2 standard deviations of the sample estimate



[†] I'm playing fast and loose here, we can't actually use the normal distribution, we use the Student's t-distribution with n-k-1 degrees of freedom. But there's no need to complicate things you don't need to know about. Look at today's class notes for more.

Interpreting Confidence Intervals

- So our confidence interval for our slope is $-3.22, -1.33), what does this mean?

❌ 95% of the time, the true effect of class size on test score will be between -3.22 and -1.33

❌ We are 95% confident that a randomly selected school district will have an effect of class size on test score between -3.22 and -1.33

❌ The effect of class size on test score is -2.28 95% of the time.

✅ We are 95% confident that in similarly constructed samples, the true effect is between -3.22 and -1.33

Hypothesis Testing

Estimation and Hypothesis Testing I

- We want to **test** if our estimates are **statistically significant** and they describe the population
  - This is the "bread and butter" of inferential statistics and the purpose of regression

Examples:

- Does reducing class size actually improve test scores?
- Do more years of education increase your wages?
- Is the gender wage gap between men and women really $0.77?

- **All modern science is built upon statistical hypothesis testing, so understand it well!**

Estimation and Hypothesis Testing II

- Note, we can test a lot of hypotheses about a lot of population parameters, e.g.
  - A population mean μ
  - **Example**: average height of adults
  - A population proportion p
  - **Example**: percent of voters who voted for Trump
  - A difference in population means $\mu_A - \mu_B$
  - **Example**: difference in average wages of men vs. women
  - A difference in population proportions $p_A - p_B$
  - **Example**: difference in percent of patients reporting symptoms of drug A vs B
  - See all the possibilities in glorious detail in today's class notes
- We will focus on hypotheses about **population regression slope** $(\hat{\beta}_1)$, i.e. the **causal effect**[†] of X on Y

[†] With a model this simple, it's almost certainly **not** causal, but this is the ultimate direction we are heading...

Null and Alternative Hypotheses I

- All scientific inquiries begin with a **null hypothesis** (H_0) that proposes a specific value of a population parameter
  - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)

- We suggest an **alternative hypothesis** (H_a), often the one we hope to verify
  - Note, can be multiple alternative hypotheses: $H_1, H_2, \ldots , H_n$

- Ask: **"Does our data (sample) give us sufficient evidence to reject H_0 in favor of H_a?"**
  - Note: **the test is *always* about $\mathbf{H_0}$**!
  - See if we have sufficient evidence to reject the status quo

Null and Alternative Hypotheses II

- Null hypothesis assigns a value (or a range) to a population parameter
  - e.g. $\beta_1 = 2$ or $\beta_1 \leq 20$
  - **Most common is** $\beta_1 = 0 \implies X$ has no effect on Y (no slope for a line)
  - Note: always an equality!

- Alternative hypothesis must mathematically *contradict* the null hypothesis
  - e.g. $\beta_1 \neq 2$ or $\beta_1 > 20$ or $\beta_1 \neq 0$
  - Note: always an inequality!

- Alternative hypotheses come in two forms:
  1. **One-sided alternative**: $\beta_1 > H_0$ or $\beta_1 < H_0$
  2. **Two-sided alternative**: $\beta_1 \neq H_0$
     - Note this means either $\beta_1 < H_0$ or $\beta_1 > H_0$

Components of a Valid Hypothesis Test

- All statistical hypothesis tests have the following components:

1. A **null hypothesis, H_0**

2. An **alternative hypothesis, H_a**

3. A **test statistic** to determine if we reject H_0 when the statistic reaches a "critical value"

    - Beyond the critical value is the "rejection region", sufficient evidence to reject H_0

4. A **conclusion** whether or not to reject H_0 in favor of H_a

Type I and Type II Errors I

- Any sample statistic (e.g. $\hat{\beta_1}$) will rarely be exactly equal to the hypothesized population parameter (e.g. β_1)

- Difference between observed statistic and true paremeter could be because:

1. Parameter is *not* the hypothesized value (H_0 is *false*)

2. Parameter is truly the hypothesized value (H_0 is *true*) but *sampling variability* gave us a different estimate

- **We cannot distinguish between these two possibilities with any certainty**

Type I and Type II Errors II

- We can interpret our estimates probabilistically as commiting one of two types of error:

1. **Type I error (false positive)**: rejecting H_0 when it is in fact true
    - Believing we found an important result when there is truly no relationship
2. **Type II error (false negative)**: failing to reject H_0 when it is in fact false
    - Believing we found nothing when there was truly a relationship to find

Type I and Type II Errors III

		Truth	
		Null is True	**Null is False**
Judgment	**Reject Null**	TYPE I ERROR	CORRECT
		(False +)	(True +)
	Don't Reject Null	CORRECT	TYPE II ERROR
		(True -)	(False -)

- Depending on context, committing one type of error may be more serious than the other

Type I and Type II Errors IV

		Truth	
		Defendant is Innocent	**Defendant is Guilty**
Judgment	**Convict**	TYPE I ERROR	CORRECT
		(False +)	(True +)
	Acquit	CORRECT	TYPE II ERROR
		(True -)	(False -)

- Anglo-American common law *presumes* defendant is innocent: H_0
- Jury judges whether the evidence presented against the defendant is plausible *assuming the defendant were in fact innocent*
- If highly improbable: sufficient evidence to reject H_0 and convict

Type I and Type II Errors V

William Blackstone

(1723-1780)

> "It is better that ten guilty persons escape than that one innocent suffer."

- Type I error is worse than a Type II error in law!

Blackstone, William, 1765-1770, *Commentaries on the Laws of England*

Type I and Type II Errors VI

Significance Level, α, and Confidence Level $1 - \alpha$

- The **significance level, α**, is the probability of a **Type I error**

$$\alpha = P(\text{Reject } H_0 | H_0 \text{ is true})$$

- The **confidence level** is defined as $(1 - \alpha)$
  - Specify *in advance* an α-level (0.10, 0.05, 0.01) with associated confidence level (90%, 95%, 99%)

- The probability of a **Type II error** is defined as β:

$$\beta = P(\text{Don't reject } H_0 | H_0 \text{ is false})$$

α and β

		Truth	
		Null is True	**Null is False**
Judgment	**Reject Null**	TYPE I ERROR	CORRECT
		α	(1-β)
	Don't Reject Null	CORRECT	TYPE II ERROR
		(1-α)	β

Power and p-values

- The statistical **power of the test** is $(1-\beta)$: the probability of correctly rejecting H_0 when H_0 is in fact false (e.g. not convicting an innocent person)

$$\text{Power} = 1-\beta = P(\text{Reject } H_0 | H_0 \text{ is false})$$

- The **p-value** or **significance probability** is the probability that, given the null hypothesis is true, the test statistic from a random sample will be at least as extreme as the test statistic of our sample

$$p(\delta \geq \delta_i | H_0 \text{ is true})$$

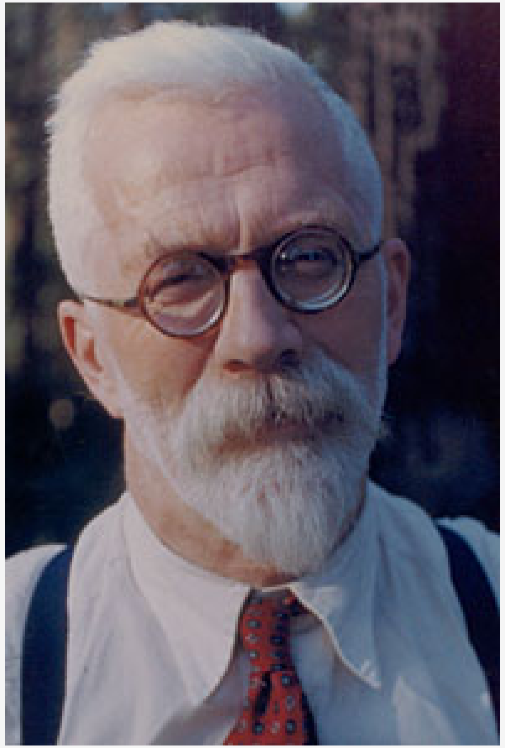
- where δ represents some test statistic
- δ_i is the test statistic we observe in our sample
- More on this in a bit

p-Values and Statistical Significance

- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)
- If $p < \alpha$: **statistically significant** evidence sufficient to *reject* H_0 in favor of H_a
  - Note this does **not** mean H_a is true! We merely have *rejected* H_0!
- If $p \geq \alpha$: *insufficient* evidence to reject H_0
  - Note this does **not** mean H_0 is true! We merely have *failed* to *reject* H_0!

Digression: p-Values and the Philosophy of Science

Hypothesis Testing and the Philosophy of Science I



Sir Ronald A. Fisher

(1890—1962)

> "The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."

1931, *The Design of Experiments*

Hypothesis Testing and the Philosophy of Science I

- Modern philosophy of science is largely based off of hypothesis testing and **falsifiability**, which form the "Scientific Method"†

- For something to be "scientific", it must be **falsifiable**, or at least **testable**

- Hypotheses can be *corroborated* with evidence, but always *tentative* until falsified by data in suggesting an alternative hypothesis

> **"All swans are white"** is a hypothesis rejected upon discovery of a single black swan

[1] Note: economics is a very different kind of "science" with a different methodology!

Hypothesis Testing and p-Values

- Hypothesis testing, confidence intervals, and p-values are probably the hardest thing to understand in statistics

Fivethirtyeight: Not Even Scientists Can Easily Explain P-values