2.7 — Inference for Regression

ECON 480 • Econometrics • Fall 2021

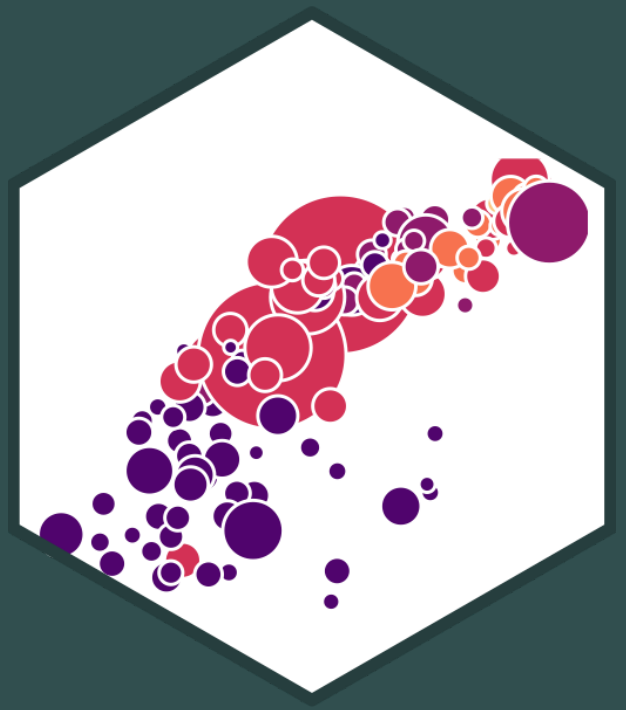
Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF21

metricsF21.classes.ryansafner.com



Outline

Hypothesis Testing

Estimation and Hypothesis Testing I

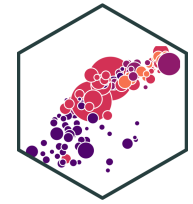
- We want to **test** if our estimates are **statistically significant** and they describe the population
  - this is the “bread and butter” of using inferential statistics

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 23%?

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

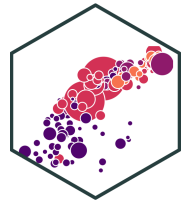
Estimation and Hypothesis Testing II



- 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
- We will focus on hypotheses about **population regression slope** $(\hat{\beta}_1)$, i.e. the **causal effect**[†] of X on Y

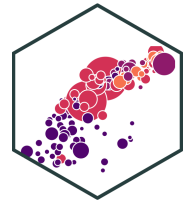
[†] With a model this simple, it's almost certainly **not** causal, but this is the ultimate direction we are heading...

Null and Alternative Hypotheses I



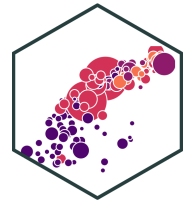
- All scientific inquiries begin with a **null hypothesis** (H_0) that proposes a specific value of a population parameter
  - Notation: add a subscript 0: $\beta_{1,0}$ (or μ_0, p_0, etc)
- We suggest an **alternative hypothesis** (H_a), often the one we hope to verify
  - Note, can be multiple alternative 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"

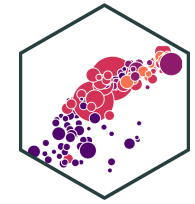
    - Beyond the critical value is the "rejection region", sufficient evidence to reject H_0

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

Type I and Type II Errors I

- Sample statistic ($\hat{\beta}_1$) will rarely be exactly equal to the hypothesized parameter (β_1)
- Difference between observed statistic and true parameter could be because:
- **Parameter is *not* the hypothesized value**
  - H_0 is *false*
- **Parameter is truly hypothesized value but *sampling variability* gave us a different estimate**
  - H_0 is *true*
- **We cannot distinguish between these two possibilities with any certainty**

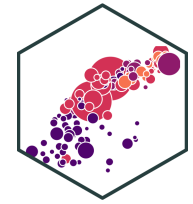
Type I and Type II Errors II



- 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

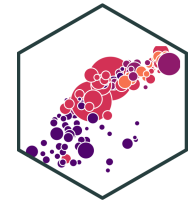
Type I and Type II Errors III



		Truth	
		Null is True	**Null is False**
Judgment	**Reject Null**	**Type I Error** (False Positive)	CORRECT (True Positive)
	Don't Reject Null	CORRECT (True Negative)	**Type II Error** (False Negative)

- 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** (False Positive)	CORRECT (True Positive)
	Acquit	CORRECT (True Negative)	**Type II Error** (False Negative)

- 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 (beyond a "reasonable doubt"): sufficient evidence to reject H_0 and convict

Type I and Type II Errors V

William Blackstone

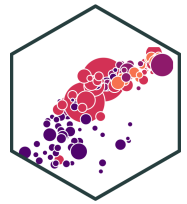
(1723-1780)

> "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



Type I and Type II Errors VI

simine vazire
@siminevazire

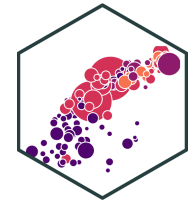
Type I error: You think you're on mute but you're not

Type II error: You think they can hear you but you're on mute

Or did I get it backwards?

11:47 PM · 2020-09-11 · Twitter for iPhone

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



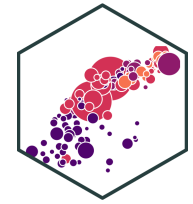
- The **significance level, α**, is the probability of a **Type I error**

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

- 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:

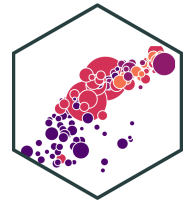
$$\beta = P(\text{Don't reject } H_0 | H_0 \text{ is false})$$

α and β



		Truth	
		Null is True	**Null is False**
Judgment	**Reject Null**	**Type I Error** (α)	CORRECT $(1-\beta)$
	Don't Reject Null	CORRECT $(1-\alpha)$	**Type II Error** (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. convicting a guilty person)

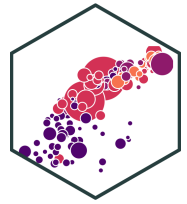
$$\text{Power} = 1 - \beta = P(\text{Reject } H_0 | H_0 \text{ is false})$$

- The **p-value** or **significance probability** is the probability that, if the null hypothesis were true, the test statistic from any sample will be *at least as extreme* as the test statistic from *our* sample

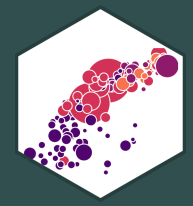
$$p(\delta \geq \delta_i | H_0 \text{ is true})$$

- 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

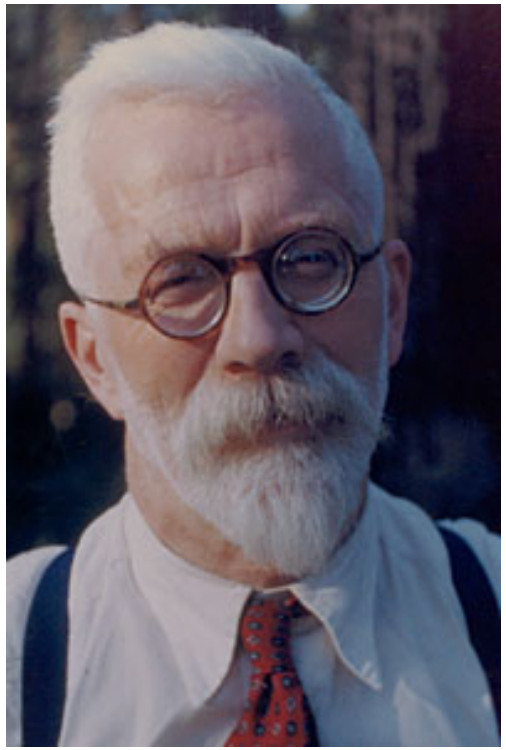


- After running our test, we need to make a *decision* between the competing hypotheses
- Compare p-value with *pre-determined* α (commonly, $\alpha = 0.05$, 95% confidence level)
- If $p < \alpha$: **statistically significant** evidence sufficient to *reject* H_0 in favor of H_a
  - Note this does **not**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

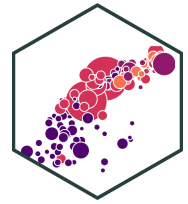
† Note: economics is a very different kind of "science" with a different methodology!

Hypothesis Testing and p-Values

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

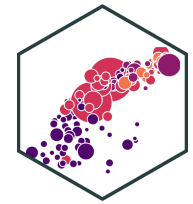
Fivethirtyeight: Not Even Scientists Can Easily Explain P-values

Hypothesis Testing: Which Test? I

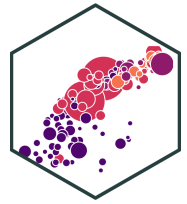


- Rigorous course on statistics (ECMG 212 or **MATH 112**) will spend weeks going through different types of tests:
  - Sample mean; difference of means
  - Proportion; difference of proportions
  - Z-test vs t-test
  - 1 sample vs. 2 samples
  - χ^2 test

Hypothesis Testing: Which Test? II



There is Only One Test



- Fortunately, some clever statisticians realized "there is only one test" and some built a nice `R` package called `infer`

1. **Calculate** a statistic, δ_i[†], from a sample of data

2. **Simulate** a world where δ is null (H_0)

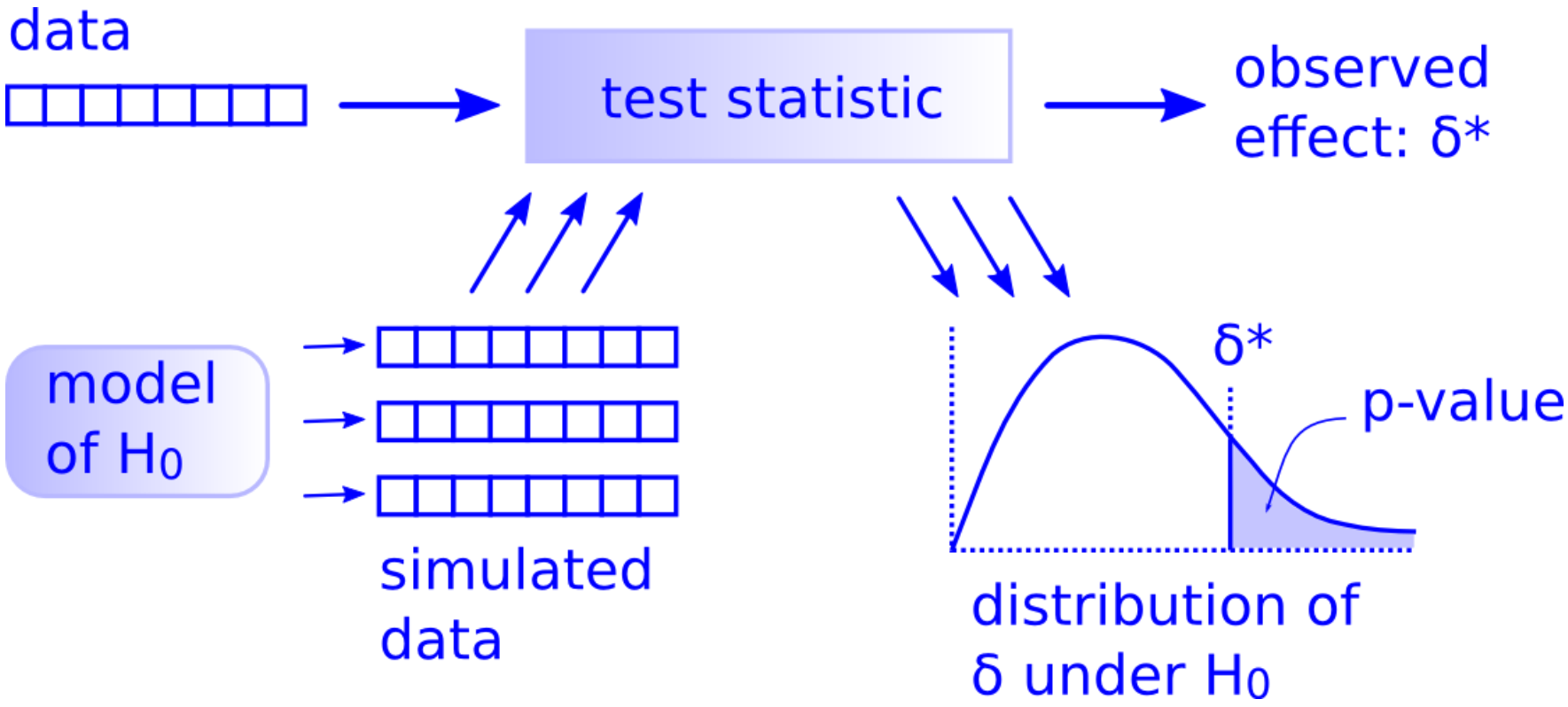
3. **Examine** the distribution of δ across the null world

4. **Calculate** the probability that δ_i could exist in the null world

5. **Decide** if δ_i is statistically significant

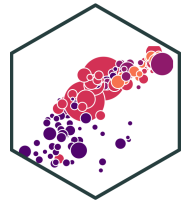
[†] δ can stand in for any test-statistic in any hypothesis test! For our purposes, δ is the slope of our regression sample, $\hat{\beta}_1$.

Elements of a Hypothesis Test



Alan Downey: “There is still only one test”

Hypothesis Testing with the infer Package I



- R naturally runs the following hypothesis test on any regression as part of `lm()`:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

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

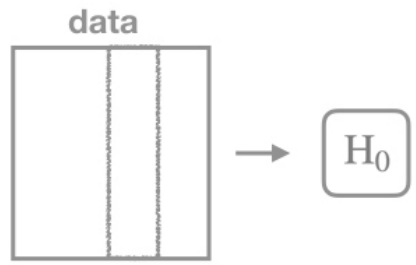
1. `specify()` a model
2. `hypothesize()` the null
3. `generate()` simulations of the null world
4. `calculate()` the p-value
5. `visualize()` with a histogram (optional)

Hypothesis Testing with the infer Package II

Hypothesis Testing with the infer Package II

`specify()`

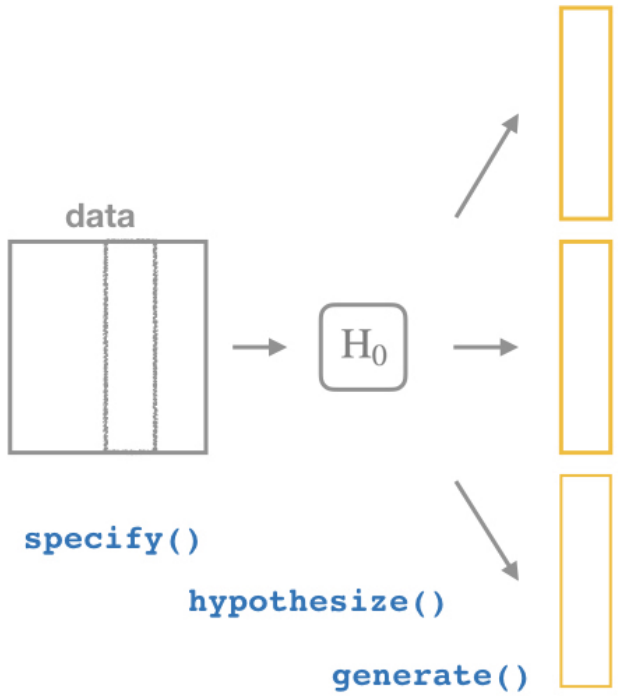
Hypothesis Testing with the infer Package II



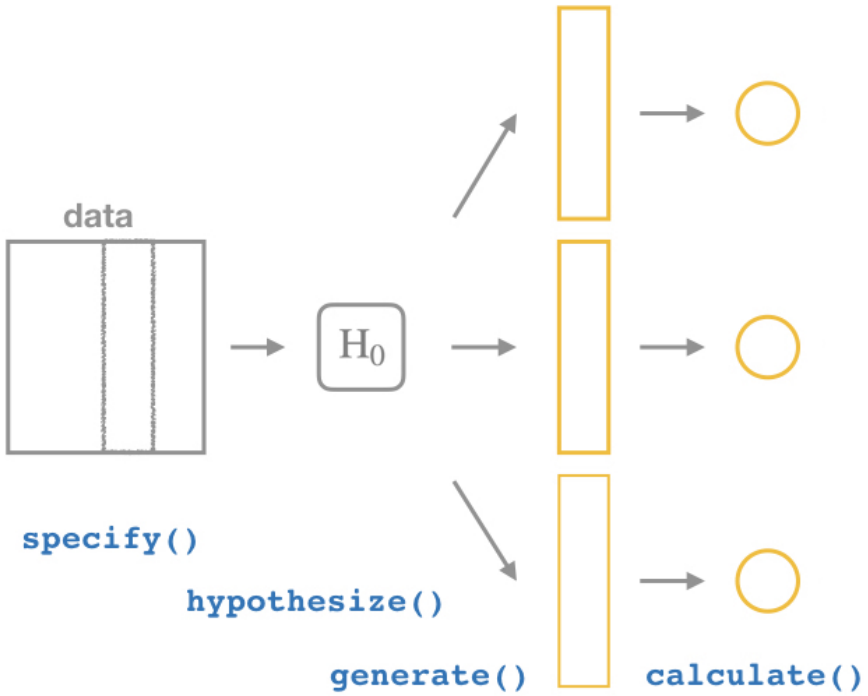
`specify()`

`hypothesize()`

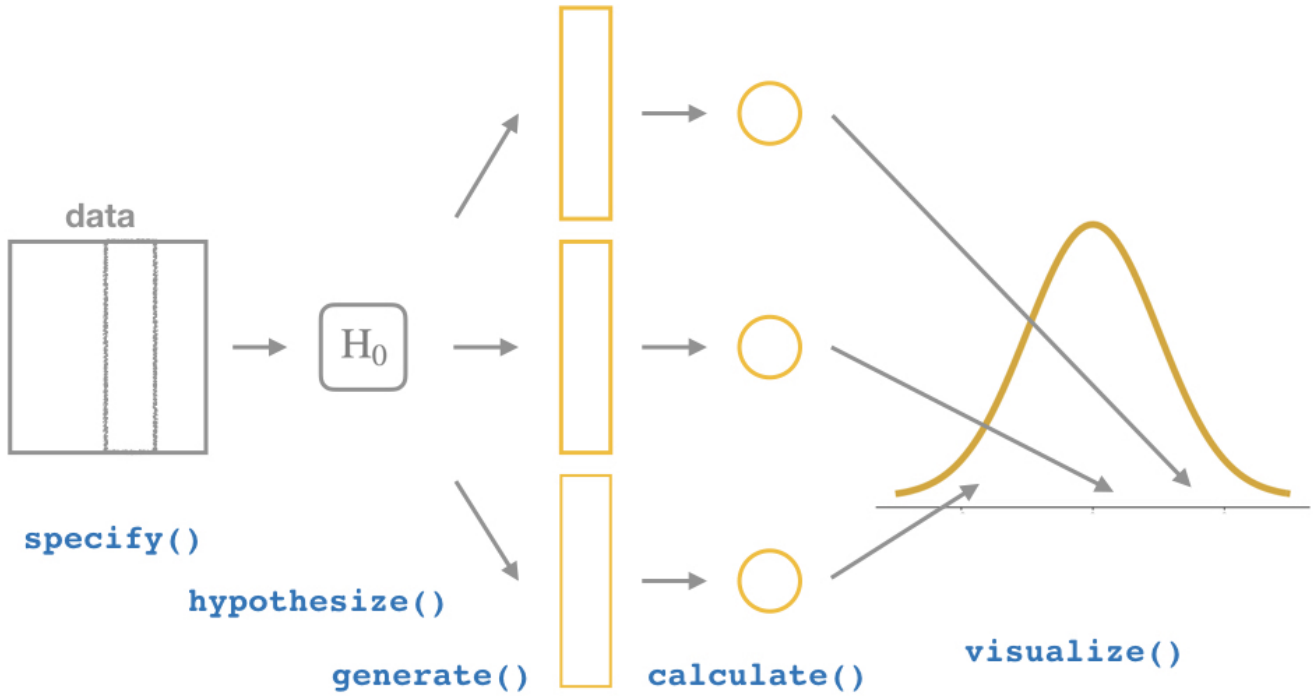
Hypothesis Testing with the infer Package II



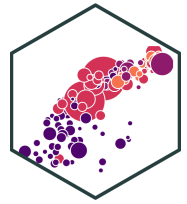
Hypothesis Testing with the infer Package II



Hypothesis Testing with the infer Package II

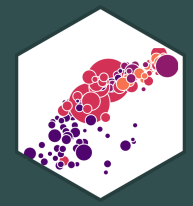


Classical Inference: Critical Values of Test Statistic



- **Test statistic** (δ): measures **how far what we observed in our sample** $(\hat{\beta}_1)$ **is from what we would expect if the null hypothesis were true** $(\beta_1 = 0)$
  - Calculated from a sampling distribution of the estimator (i.e. $\hat{\beta}_1$)
  - In econometrics, we use t-distributions which have $n - k - 1$ degrees of freedom[†]
- **Rejection region**: if the test statistic reaches a **"critical value"** of δ, then we **reject** the null hypothesis

[†] Again, see last class's appendix for more on the t-distribution. k is the number of independent variables our model has, in this case, with just one X, $k = 1$. We use two degrees of freedom to calculate $\hat{\beta}_0$ and $\hat{\beta}_1$, hence we have $n - 2$ df.

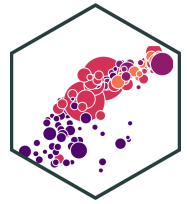


Hypothesis Testing by Simulation, with *infer*

Imagine a Null World, where H_0 is True

Our world, and a world where $\beta_1 = 0$ by assumption.

Comparing the Worlds I



- From that null world where $H_0 : \beta_1 = 0$ is true, we **simulate** another sample and calculate OLS estimators again

Our Sample

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

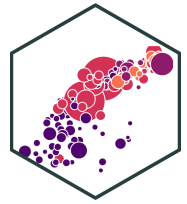
2 rows | 1-3 of 5 columns

Another Sample

term	estimate	std.error
<chr>	<dbl>	<dbl>
(Intercept)	647.8027952	9.7147718
str	0.3235038	0.4923581

2 rows | 1-3 of 5 columns

Comparing the Worlds II



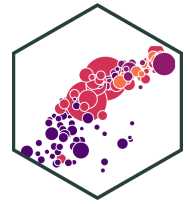
- From that null world where $H_0 : \beta_1 = 0$ is true, let's **simulate 1,000** samples and calculate slope $(\hat{\beta}_1)$ for each

sample <int>	slope <dbl>
1	-0.3027333296
2	-0.3624481355
3	0.6448518690
4	-0.0745971847
5	0.5969444290
6	0.5505335318
7	0.5927466147
8	0.0572148658
9	-0.0989989073
10	0.8043957511

1-10 of 1,000 rows

Previous **1** 2 3 4 5 6 … 100 Next

Prepping the *infer* Pipeline



- Before I show you how to do this, let's first save our estimated slope from our *actual* sample
  - We'll want this later!

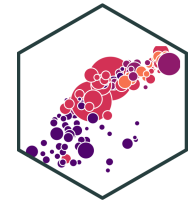
```
# save as obs_slope
sample_slope <- school_reg_tidy %>% # this is the regression tidied with broom's tidy()
  filter(term=="str") %>%
  pull(estimate)

# confirm what it is
sample_slope
```

```
## [1] -2.279808
```

The *infer* Pipeline: Specify

`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)

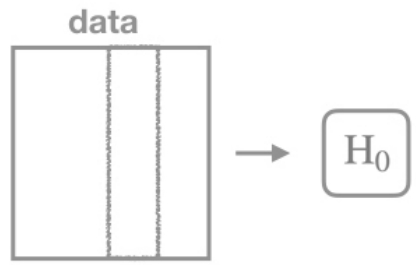
```
CASchool %>%
  specify(testscr ~ str)
```

testscr	str
<dbl>	<dbl>
690.8	17.88991
661.2	21.52466
643.6	18.69723

3 rows

- Note nothing happens yet

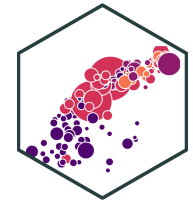
The *infer* Pipeline: Hypothesize




```
specify()
```

```
hypothesize()
```

The *infer* Pipeline: Hypothesize



Specify

Hypothesize

`%>% hypothesize(null = "independence")`

- Describe what the null hypothesis is here
- In `infer`'s language, we are hypothesizing that `str` and `testscr` are `independent` $(\beta_1 = 0)$[†]

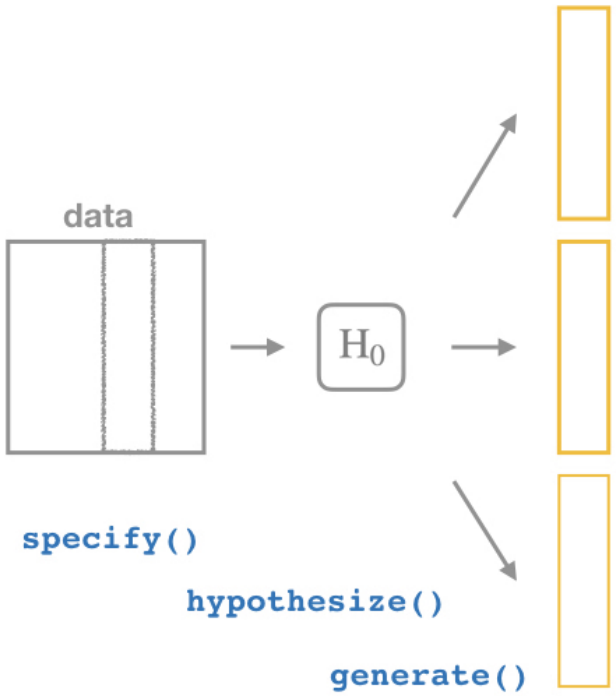
```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence")
```

testscr	str
<dbl>	<dbl>
690.8	17.88991
661.2	21.52466
643.6	18.69723

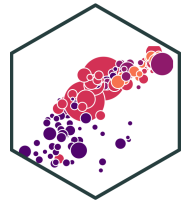
3 rows

[†] See more here about what other hypotheses you can test with `infer`.

The *infer* Pipeline: Generate I



The *infer* Pipeline: Generate I



Specify

Hypothesize

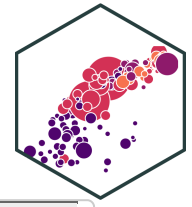
Generate

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

- Now the magic starts, as we run a number of simulated samples
- Set the number of `reps` and set the `type` equal to `"permute"`
  - we want `permutation` (not `bootstrap`!) because we are simulating a world where $\beta_1 = 0$ by construction!

```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000,
           type = "permute")
```

The *infer* Pipeline: Generate II



Specify

Hypothesize

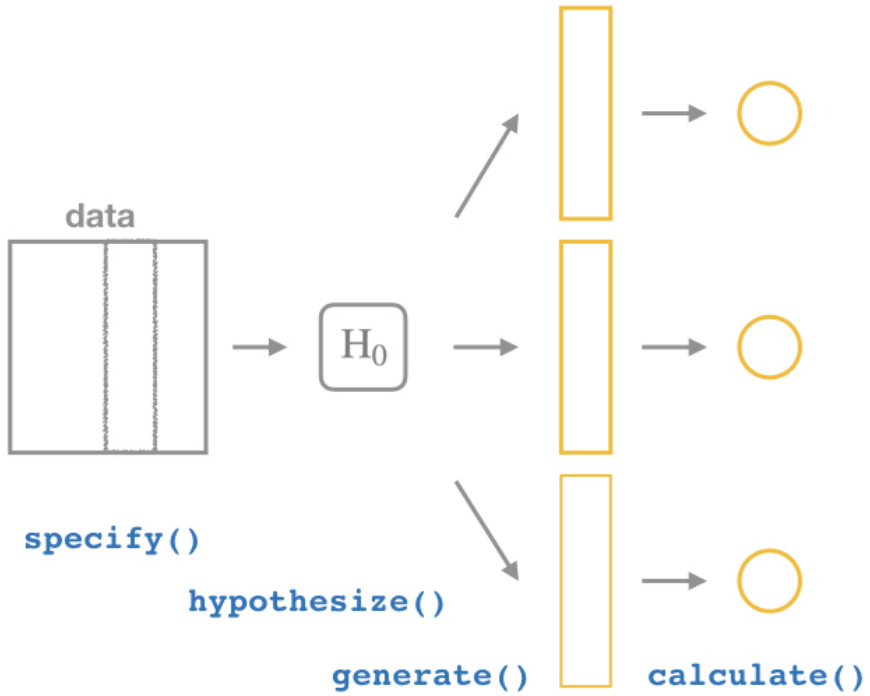
Generate

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

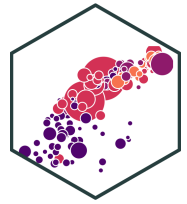
testscr <dbl>	str <dbl>	replicate <int>
693.95	17.88991	1
642.40	21.52466	1
680.45	18.69723	1
672.70	17.35714	1
666.45	18.67133	1
654.20	21.40625	1
671.95	19.50000	1
671.75	20.89412	1
624.55	19.94737	1
699.10	20.80556	1

1-10 of 10,000 rows Previous **1** 2 3 4 5 6 … 1000 Next

The *infer* Pipeline: Calculate I



The *infer* Pipeline: Calculate I



Specify

Hypothesize

Generate

Calculate

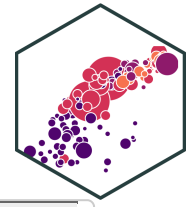
`%>% calculate(stat = "")`

- We `calculate` sample statistics for each of the 1,000 `replicate` samples
- In our case, calculate the slope, $(\hat{\beta}_1)$ for each `replicate`

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

- Other `stat`s for calculation: `"mean"`, `"median"`, `"prop"`, `"diff in means"`, `"diff in props"`, etc. (see package information)

The *infer* Pipeline: Calculate II



Specify

Hypothesize

Generate

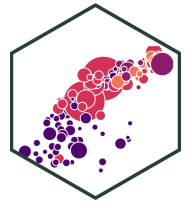
Calculate

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

replicate <int>	stat <dbl>
1	0.384783281
2	0.241700895
3	0.268799843
4	-0.189039951
5	1.215030315
6	0.511783627
7	-0.457378304
8	1.008206723
9	0.092043084
10	0.233837801

1-10 of 1,000 rows Previous **1** 2 3 4 5 6 … 100 Next

The *infer* Pipeline: Get p Value



Specify

Hypothesize

Generate

Calculate

Get p Value

```
%>% get_p_value(obs stat = "",
direction = "both")
```

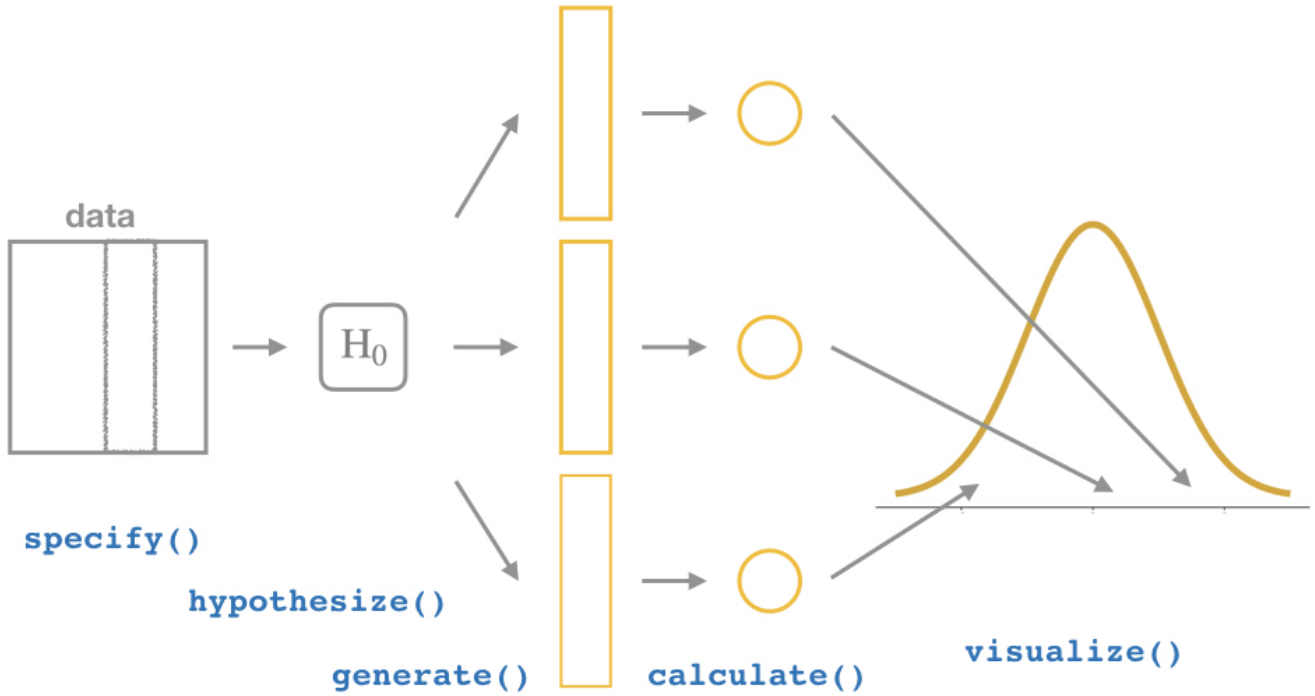
- We can calculate the **p-value**
  - the probability of seeing a value at least as large as our `sample_slope` (-2.28) in our simulated null distribution
- **Two-sided alternative** $H_a : \beta_1 \neq 0$, we double the raw p-value

```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000,
           type = "permute") %>%
  calculate(stat = "slope") %>%
  get_p_value(obs_stat = sample_slope,
              direction = "both")
```

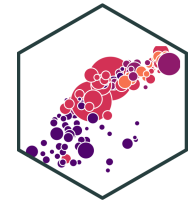
p_value
<dbl>
0

1 row

The *infer* Pipeline: Visualize I



The *infer* Pipeline: Visualize I



Specify

Hypothesize

Generate

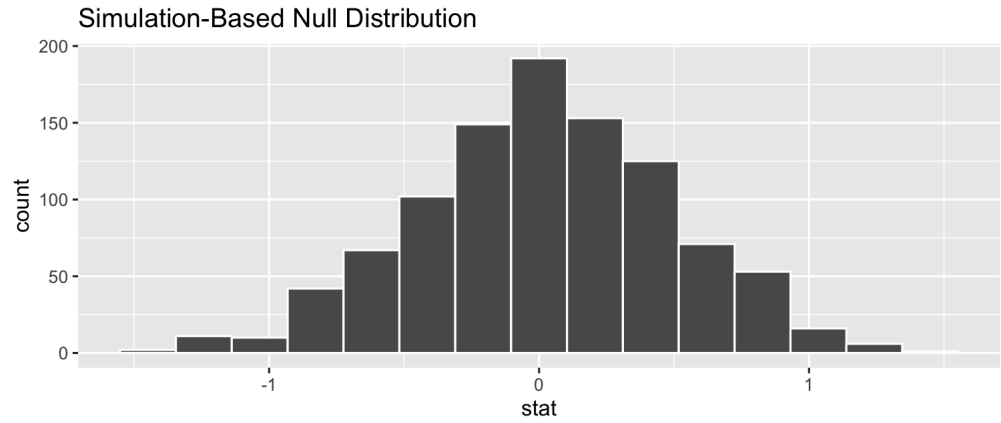
Calculate

Visualize

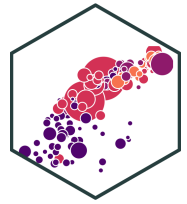
```
%>% visualize()
```

- Make a histogram of our null distribution of β_1
  - Note it is centered at $\beta_1 = 0$ because that's H_0!

```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000,
           type = "permute") %>%
  calculate(stat = "slope") %>%
  visualize()
```



The *infer* Pipeline: Visualize II



Specify

Hypothesize

Generate

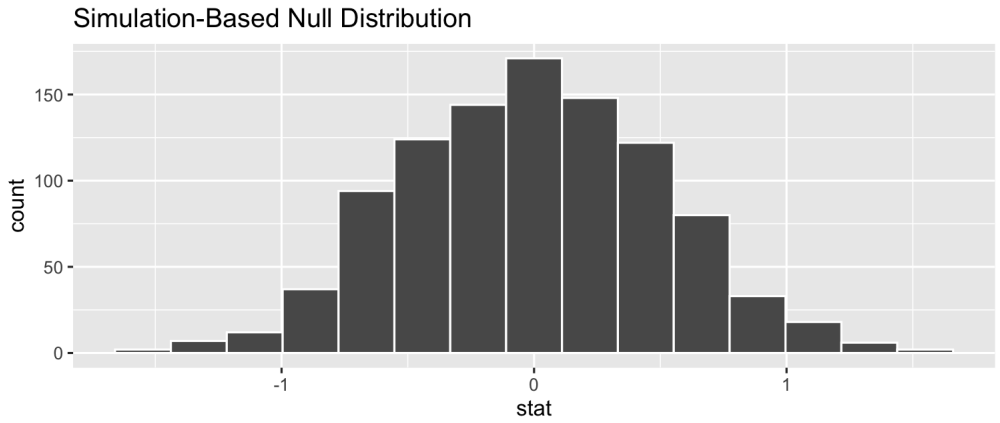
Calculate

Visualize

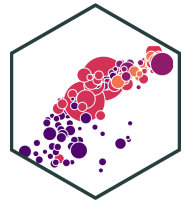
`%>% visualize()`

- Add our `sample_slope` to show our finding on the null distr.

```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000,
           type = "permute") %>%
  calculate(stat = "slope") %>%
  visualize(obs_stat = sample_slope)
```



The *infer* Pipeline: Visualize p-value



Specify

Hypothesize

Generate

Calculate

Visualize

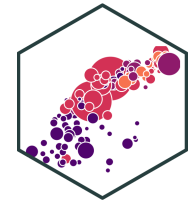
`%>%`

`visualize()+shade_p_value()`

- Add `shade_p_value()` to see what p is

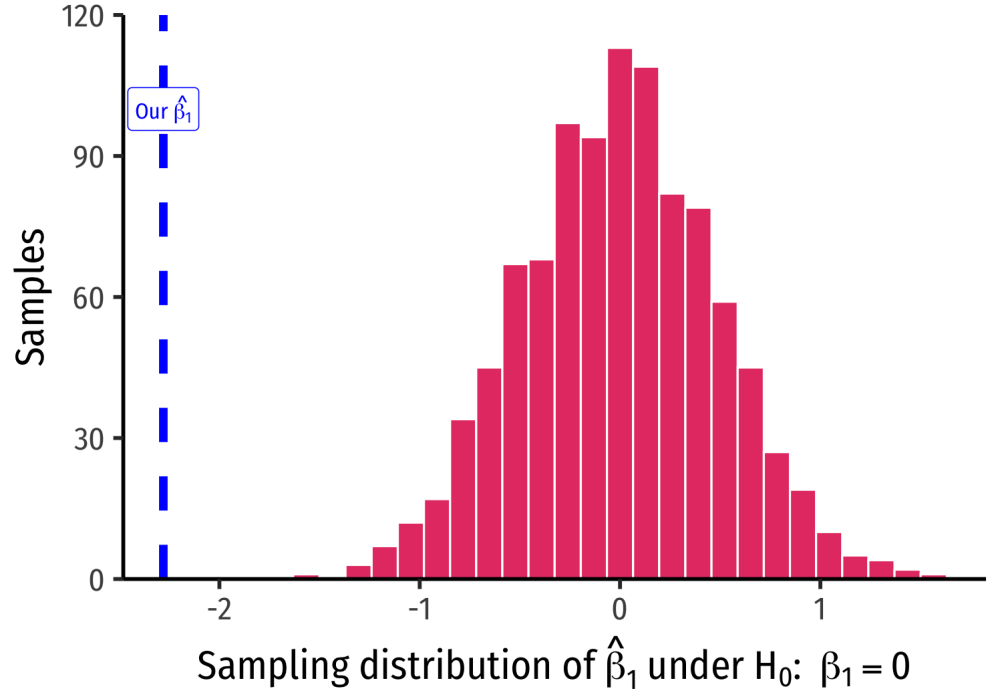
```
CASchool %>%
  specify(testscr ~ str) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000,
           type = "permute") %>%
  calculate(stat = "slope") %>%
  visualize(obs_stat = sample_slope)+
  shade_p_value(obs_stat = sample_slope,
                direction = "two_sided")
```

The *infer* Pipeline: Visualize is a Wrapper of ggplot



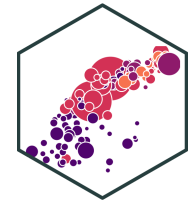
- `infer`'s `visualize()` function is just a wrapper function for `ggplot()`
  - you can take your `simulations` `tibble` and just `ggplot` a normal histogram

```
simulations %>%
  ggplot(data = .)+
  aes(x = stat)+
  geom_histogram(color="white", fill="#e64173")+
  geom_vline(xintercept = sample_slope,
             color = "blue",
             size = 2,
             linetype = "dashed")+
  annotate(geom = "label",
           x = -2.28,
           y = 100,
           label = expression(paste("Our ", hat(beta[1]))),
           color = "blue")+
  scale_y_continuous(lim=c(0,120),
                     expand = c(0,0))+
  labs(x = expression(paste("Sampling distribution of ", hat(
       y = "Samples")+
    theme_classic(base_family = "Fira Sans Condensed",
           base_size=20)
```



Classical Statistical Inference (What R Calculates)

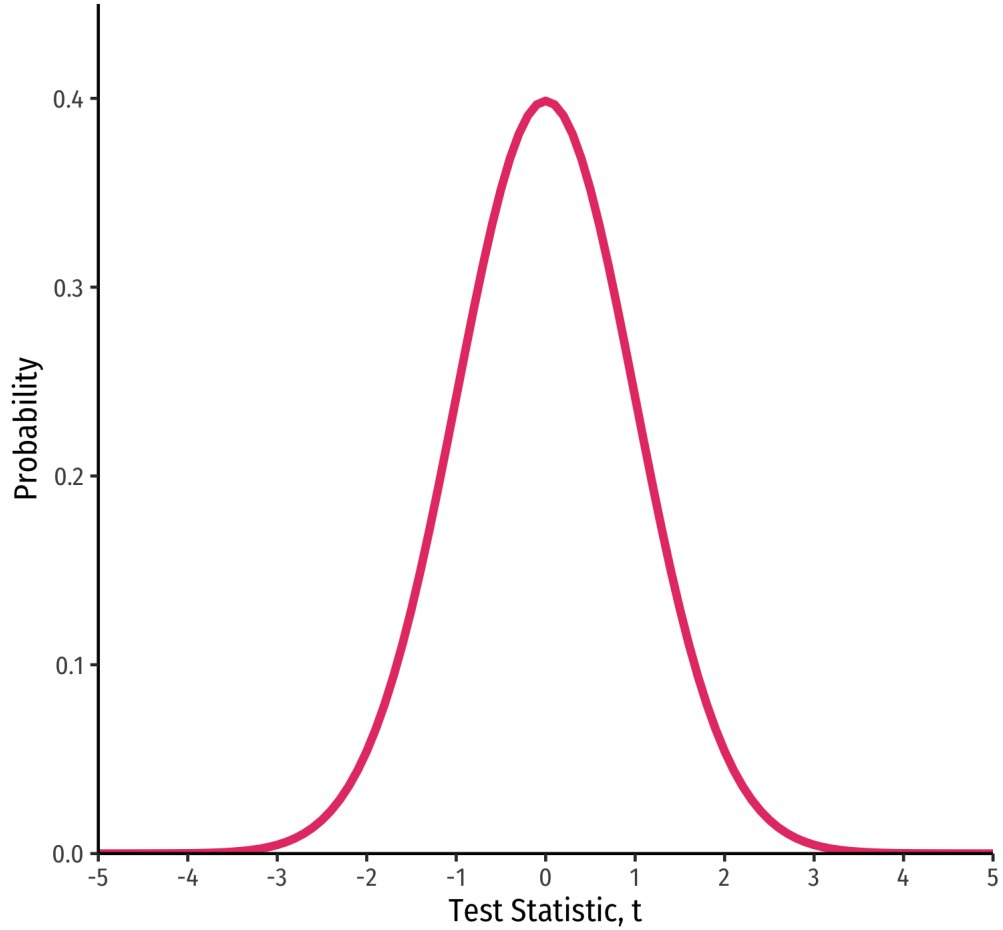
What R Does: Classical Statistical Inference I



- R does things the old-fashioned way, using a *theoretical* null distribution instead of *simulating* one
- A t-distribution with $n - k - 1$ df[†]
- Calculate a t-statistic for $\hat{\beta}_1$:

$$\text{test statistic} = \frac{\text{estimate} - \text{null hypothesis}}{\text{standard error of estimate}}$$

[†] k is the number of X variables.



t measures number of std. devs our $\hat{\beta}_1$ is from $E[\hat{\beta}_1]$ if H_0 were True

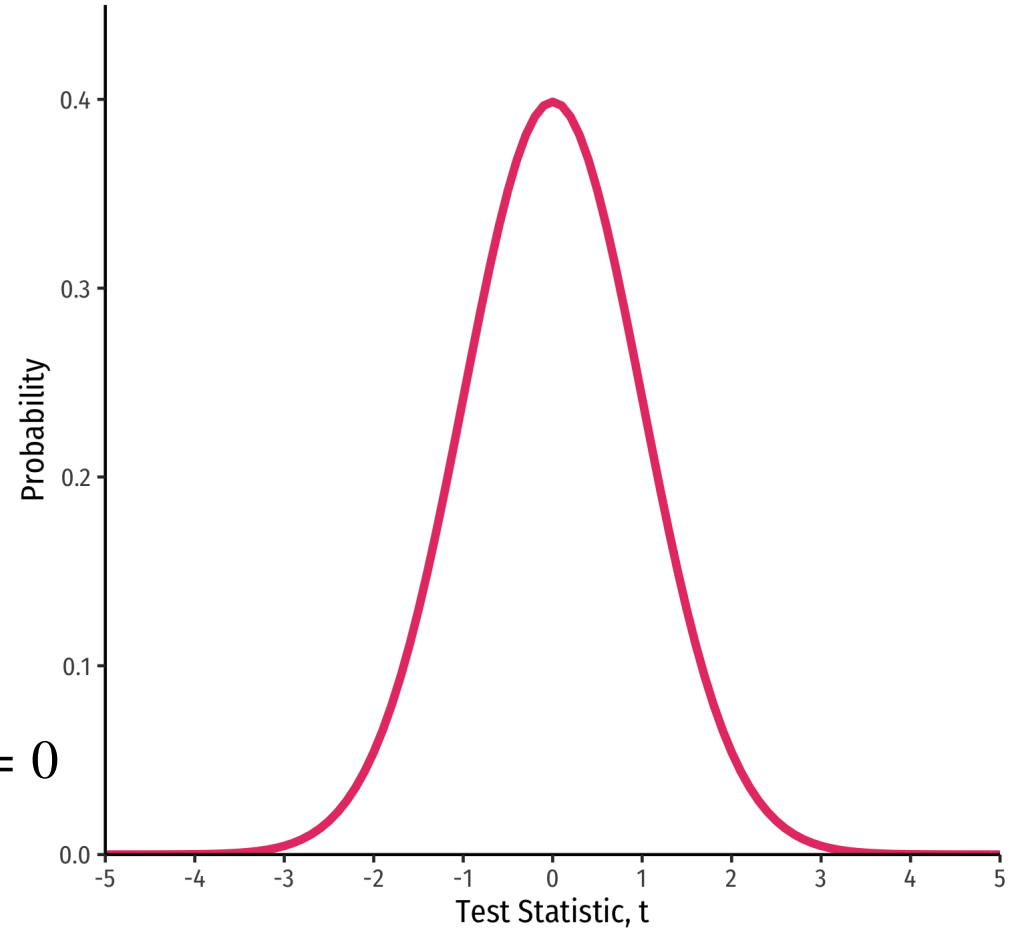
What R Does: Classical Statistical Inference II

$$\text{test statistic} = \frac{\text{estimate} - \text{null hypothesis}}{\text{standard error of estimate}}$$

- t same interpretation as Z: number of std. dev. away from the sampling distribution's expected value $E[\hat{\beta}_1]$[†] (if H_0 were true)
- Compares to a **critical value** of t^* (pre-determined by α-level & $n - k - 1$ df)
  - For 95% confidence, $\alpha = 0.05, t^* \approx 2$[‡]

[†] The expected value is 0, because our null hypothesis was $\beta_1 = 0$

[‡] Again, the **68-95-99.7%** empirical rule!

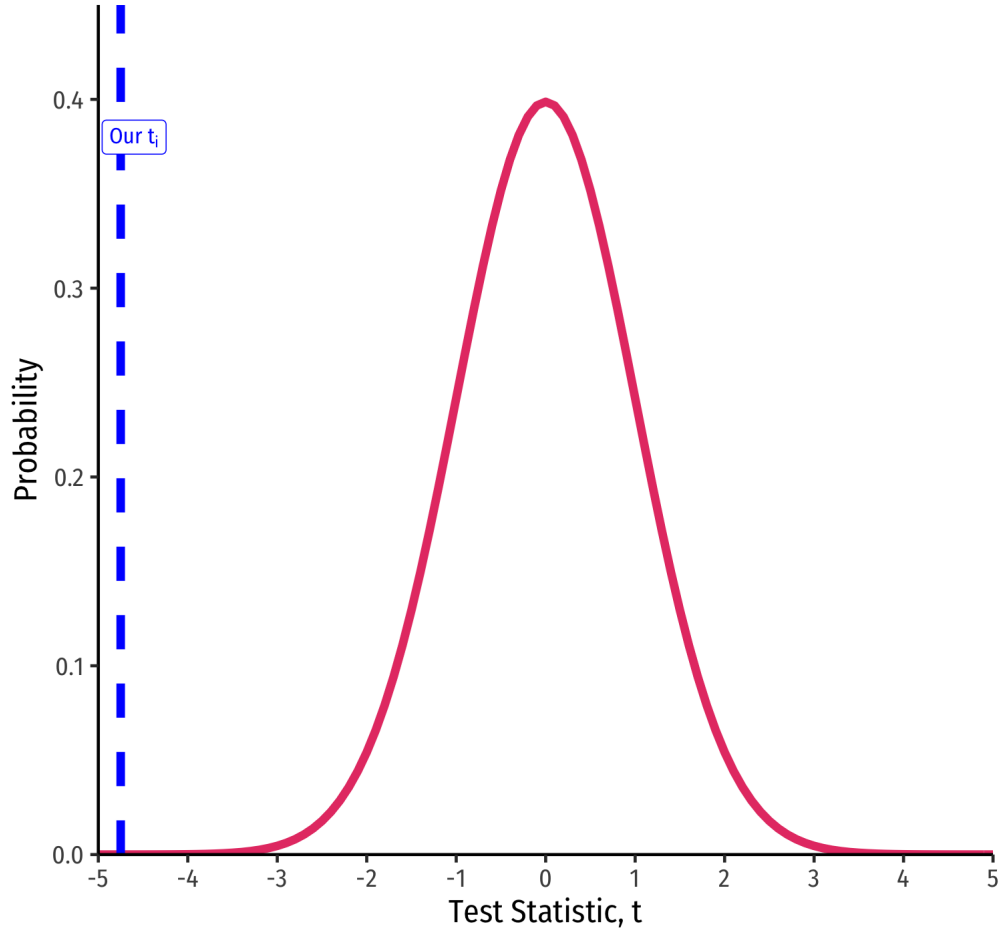


t measures number of std. devs our $\hat{\beta}_1$ is from $E[\hat{\beta}_1]$ if H_0 were True

What R Does: Classical Statistical Inference III

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$
$$t = \frac{-2.28 - 0}{0.48}$$
$$t = -4.75$$

- Our sample slope $\hat{\beta}_1$ is **4.75 standard deviations below** the expected value $E[\hat{\beta}_1]$ (i.e. 0) if H_0 were true

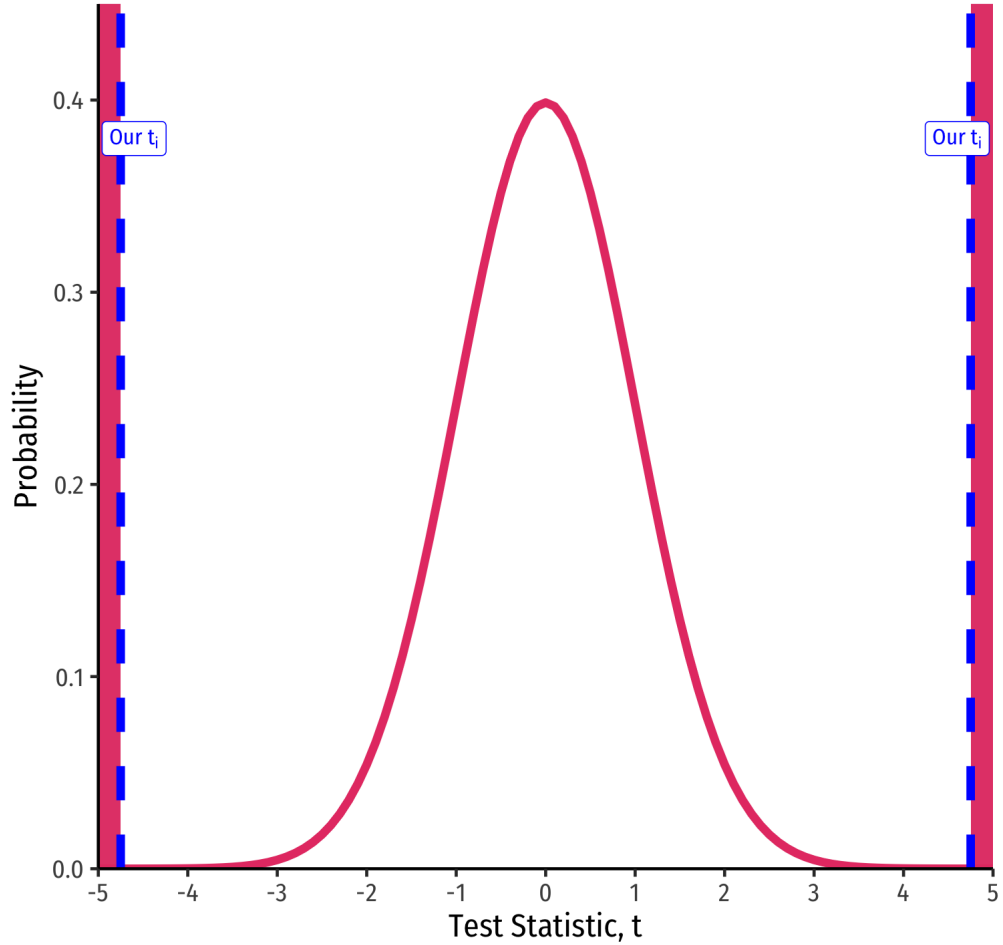


t measures number of std. devs our $\hat{\beta}_1$ is from $E[\hat{\beta}_1]$ if H_0 were True

What R Does: Classical Statistical Inference III

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$
$$t = \frac{-2.28 - 0}{0.48}$$
$$t = -4.75$$

- **p-value**: prob. of a test statistic at least as large (in magnitude) as ours if the null hypothesis were true
    - Continuous distribution implies we need probability of area *beyond* our value
    - p-value is **2-sided** for $H_a : \beta_1 \neq 0$
- $2 \times p(t_{418} > |-4.75|) = 0.0000028$



p-value is probability in the tails beyond our t

1-Sided Tests & p-values

$H_a : \beta_1 < 0$

p-value: $p(t \leq t_i)$

p-value is probability in the tail(s) beyond our test statistic, t_i of our sample slope $\hat{\beta}_1$

$H_a : \beta_1 > 0$

p-value: $p(t \geq t_i)$

p-value is probability in the tail(s) beyond our test statistic, t_i of our sample slope $\hat{\beta}_1$

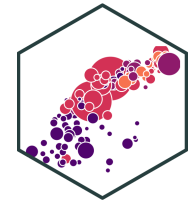
2-Sided Tests and p-values

$H_a : \beta_1 \neq 0$

p-value: $2 \times p(t \geq |t_i|)$

p-value is probability in the tail(s) beyond our test statistic, t_i of our sample slope $\hat{\beta}_1$

Calculating p-values in R

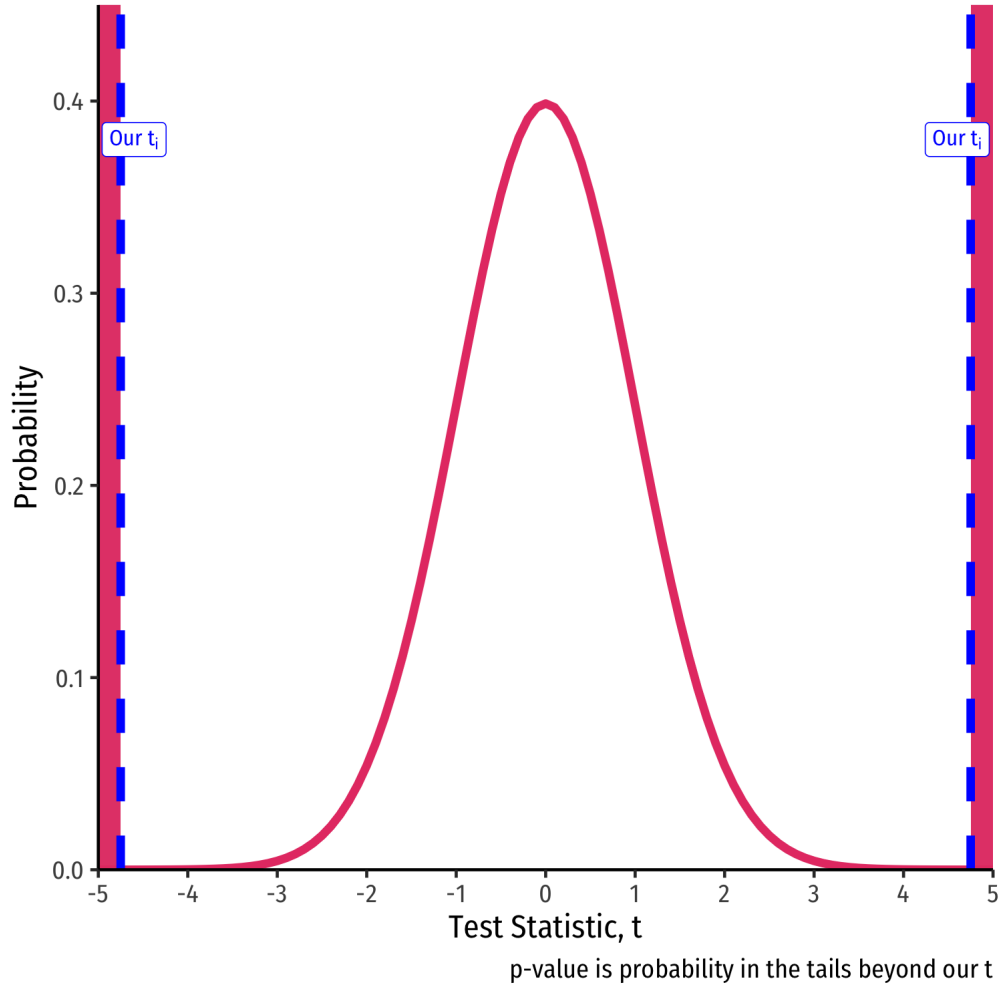


- `pt()` calculates `p`robabilities on a `t` distribution with arguments:
  - the t-score
  - `df =` the degrees of freedom
  - `lower.tail =`
    - `TRUE` if looking at area to *LEFT* of value
    - `FALSE` if looking at area to *RIGHT* of value

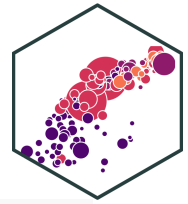
```
2 * pt(4.75, # I'll double the right tail
       df = 418,
       lower.tail = F) # right tail
```

```
## [1] 2.800692e-06
```

- $2 \times p(t_{418} > |-4.75|) = 0.0000028$



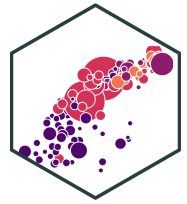
Hypothesis Tests in Regression Output I



```
summary(school_reg)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Hypothesis Tests in Regression Output II



- In `broom`'s `tidy()` (with confidence intervals)

```
tidy(school_reg, conf.int=TRUE)
```

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

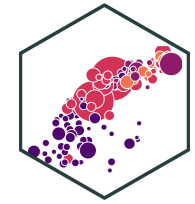
- p-value on `str` is 0.00000278.

Conclusions

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_a \neq 0$$

- Because the hypothesis test's p-value $< \alpha$ (0.05)...
- **We have sufficient evidence to reject H_0 in favor of our alternative hypothesis. Our sample suggests that there *is a relationship* between class size and test scores**.
- Using the confidence intervals:
- **We are 95% confident that, from similarly constructed samples, the true marginal effect of class size on test scores is between -3.22 and -1.34.**

Hypothesis Testing vs. Confidence Intervals



- Confidence intervals are all *two-sided* by nature

$$CI_{0.95} = \left(\left[\hat{\beta}_1 - \underbrace{2 \times se(\hat{\beta}_1)}_{MOE} \right], \left[\hat{\beta}_1 + \underbrace{2 \times se(\hat{\beta}_1)}_{MOE} \right] \right)$$

- Hypothesis test (t-test) of $H_0 : \beta_1 = 0$ computes a t-value of[1]

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

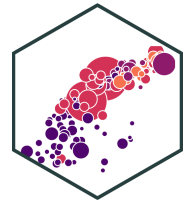
and $p < 0.05$ when $t \geq 2$ (approximately)

[1] Since our null hypothesis is that $\beta_{1,0} = 0$, the test statistic simplifies to this neat fraction.

- **If our confidence interval contains the H_0 value (i.e. 0, for our test), then *we fail to reject* H_0.**

The Use and Abuse of p-values

Common Misconceptions about p-values



- So how do we interpret p again?

❌ **p is the probability that the alternative hypothesis is false**

- We can never *prove* an alternative hypothesis, only tentatively reject a null hypothesis

❌ **p is the probability that the null hypothesis is true**

- We're not *proving* the H_0 is false, only saying that it's very unlikely that if H_0 were true, we'd obtain a slope as rare as our sample's slope

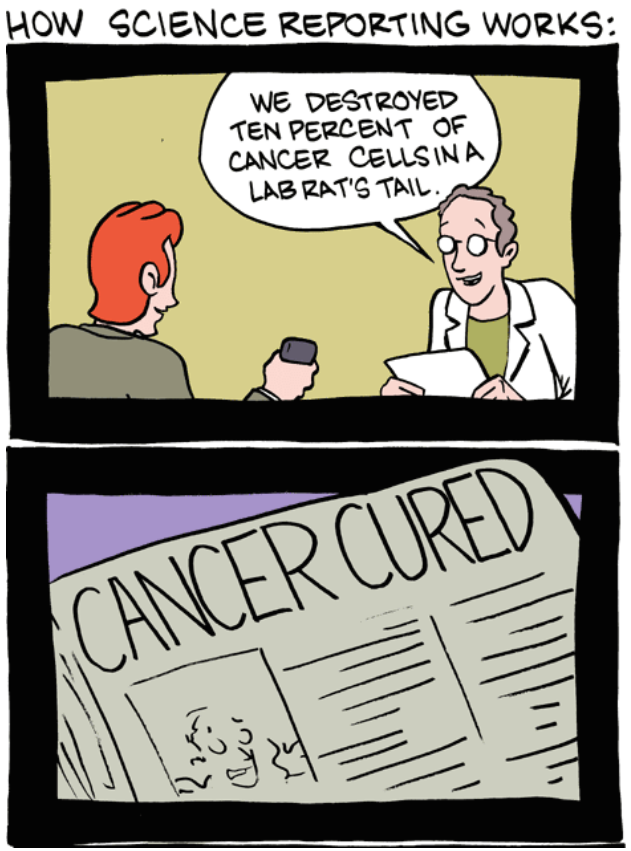
❌ **p is the probability that our observed effects were produced purely by random chance**

- p is computed under a specific model (think about our null world) that *assumes* H_0 is true

❌ **p tells us how significant our finding is**

- p tells us nothing about the *size* or the *real world significance* of any effect deemed "statistically significant"
- it only tells us that the slope is statistically significantly different from 0 (if H_0 is $\beta_1 = 0$)

Abusing p-Values I



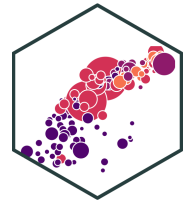
Source: SMBC

Abusing p-Values II

> “The widespread use of 'statistical significance' (generally interpreted as $(p \leq 0.05)$ as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.”

Wasserstein, Ronald L. and Nicole A. Lazar, (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician* 30(2): 129-133

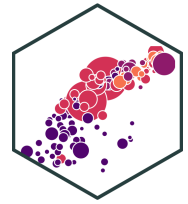
Abusing p-Values II



> “No economist has achieved scientific success as a result of a statistically significant coefficient. Massed observations, clever common sense, elegant theorems, new policies, sagacious economic reasoning, historical perspective, relevant accounting, these have all led to scientific success. Statistical significance has not.”

McCloskey, Dierdre N and Stephen Ziliak, 1996, *The Cult of Statistical Significance*, p. 112)

p-value Clarification

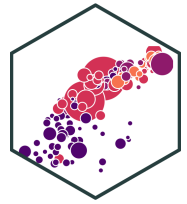


- Again, **p-value is the probability that, if the null hypothesis were true, we obtain (by pure random chance) a test statistic at least as extreme as the one we estimated for our sample**

- A low p-value means either (and we can't distinguish which):

    1. H_0 is true and a highly improbable event has occurred OR
    2. H_0 is false

Significance In Regression Tables



	Test Score
Intercept	698.93 ***
	(9.47)
STR	-2.28 ***
	(0.48)
N	420
R-Squared	0.05
SER	18.58

*** p < 0.001; ** p < 0.01; * p < 0.05.

- Statistical significance is shown by asterisks, common (but not always!) standard:
  - 1 asterisk: significant at $\alpha = 0.10$
  - 2 asterisks: significant at $\alpha = 0.05$
  - 3 asterisks: significant at $\alpha = 0.01$
- Rare, but sometimes regression tables include p-values for estimates