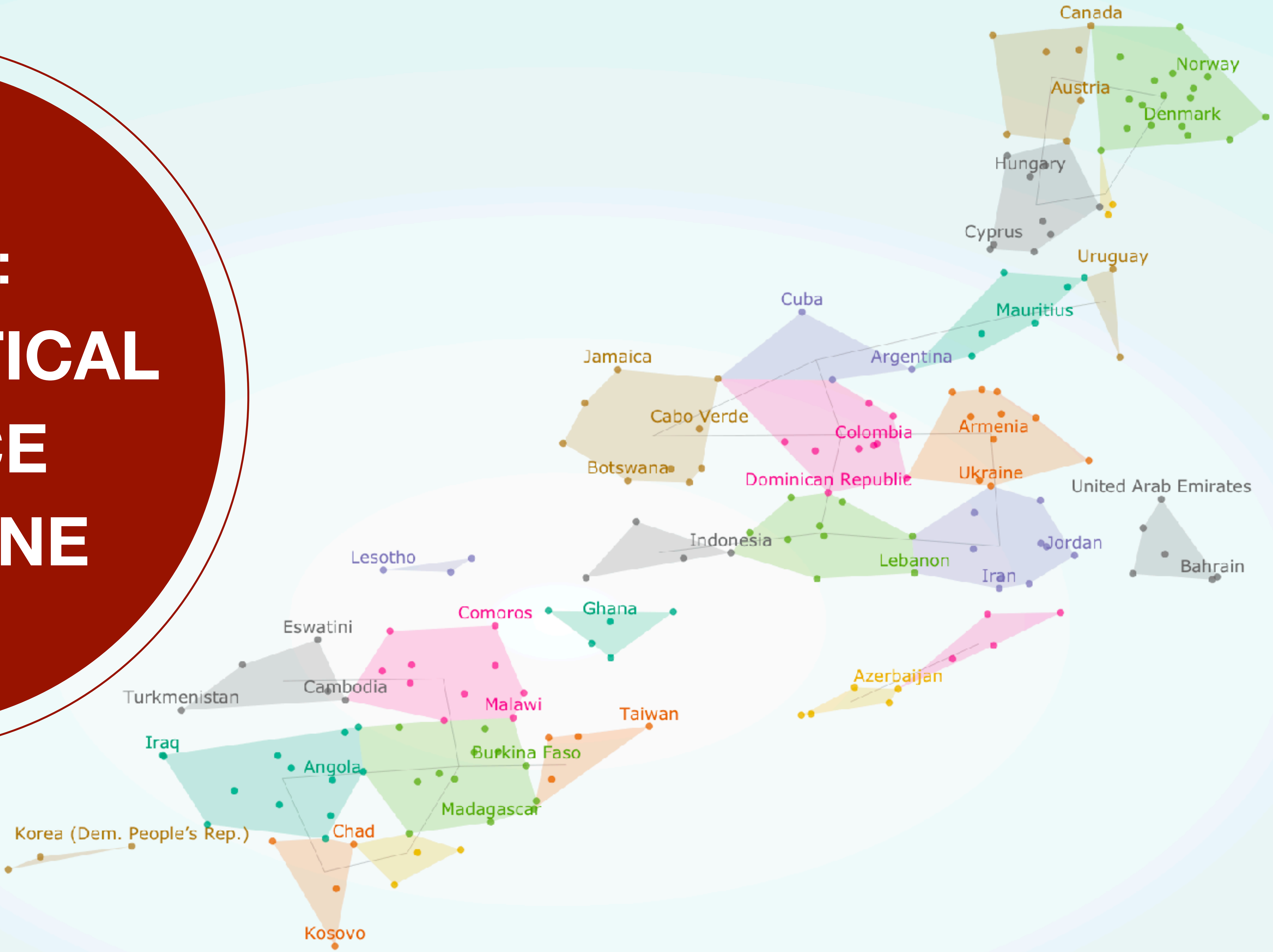
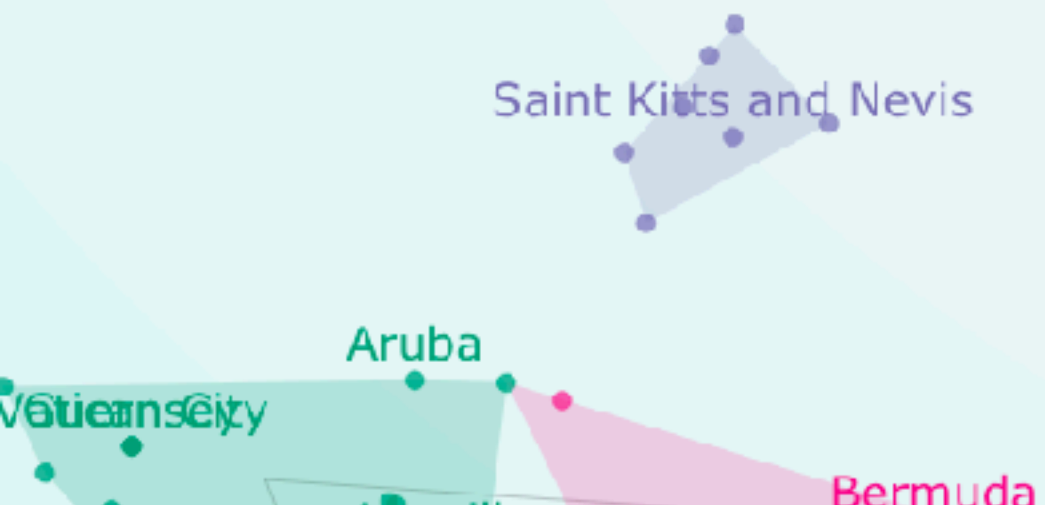
Poli-301:

The Political Science Discipline





TODAY'S AGENDA

1. Hypothesis Testing
2. P-values
3. Wrap up

Housekeeping

Thursday is review for final

Answer whatever questions/doubts you have

Finish updating website tonight

Do people with kids have more affairs than people without kids?

H1: adults with children are **more likely** to have affairs than adults without children

VS.

H0: adults with children have affairs **at same rate** as adults without children

Which hypothesis corresponds better to the data?

Our hypothesis (H1)

Null hypothesis (H0)

Coefficient on children > 0

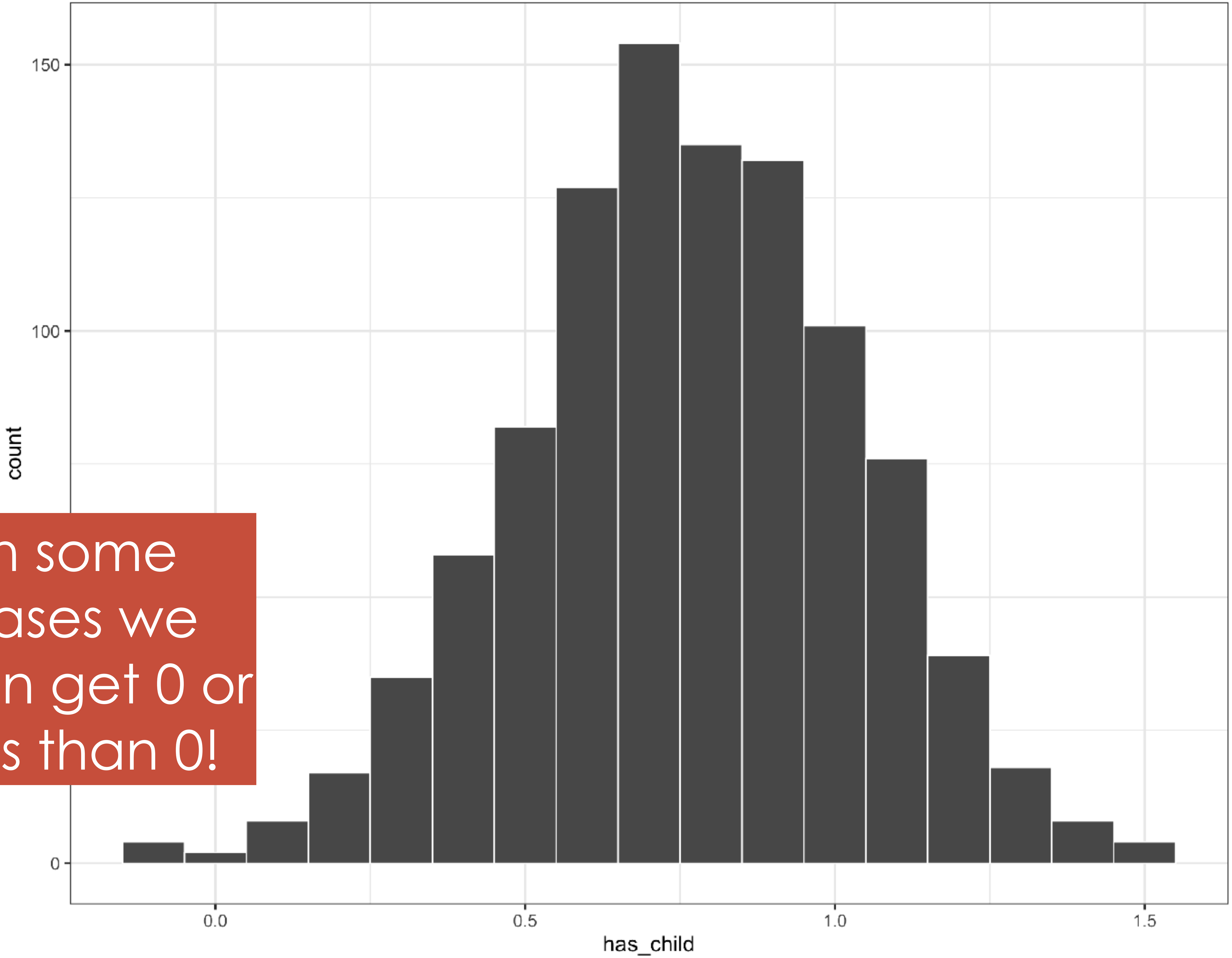
Coefficient on children = 0

```
lm(affairs ~ children, data = Affairs)
```

	term	estimate
1	intercept	0.912
2	childrenyes	0.76

Remember we have a **sample** and results on each one will be different

```
boot_affairs = Affairs %>%
specify(formula = affairs ~ children) %>%
generate(reps = 1000, type = "bootstrap") %>%
calculate(stat = "diff in means", order =
c("yes", "no"))
```



In some cases we even get 0 or less than 0!

Our hypothesis (H1)

Null hypothesis (H0)

Coefficient on children > 0

Coefficient on children = 0

Seems H1 is more likely than H0 because, more often than not, coefficient on children > 0

But how do we decide?

Science is like the courts

Presume no effect of X on Y (the **null hypothesis**)

You have burden of proving **there is an effect**

Decide if there is an effect based on **amount of evidence**

Never prove that null is true; we try and fail to reject the null

Compare what we actually observed (the coefficient from our model) against **what we might observe by chance**

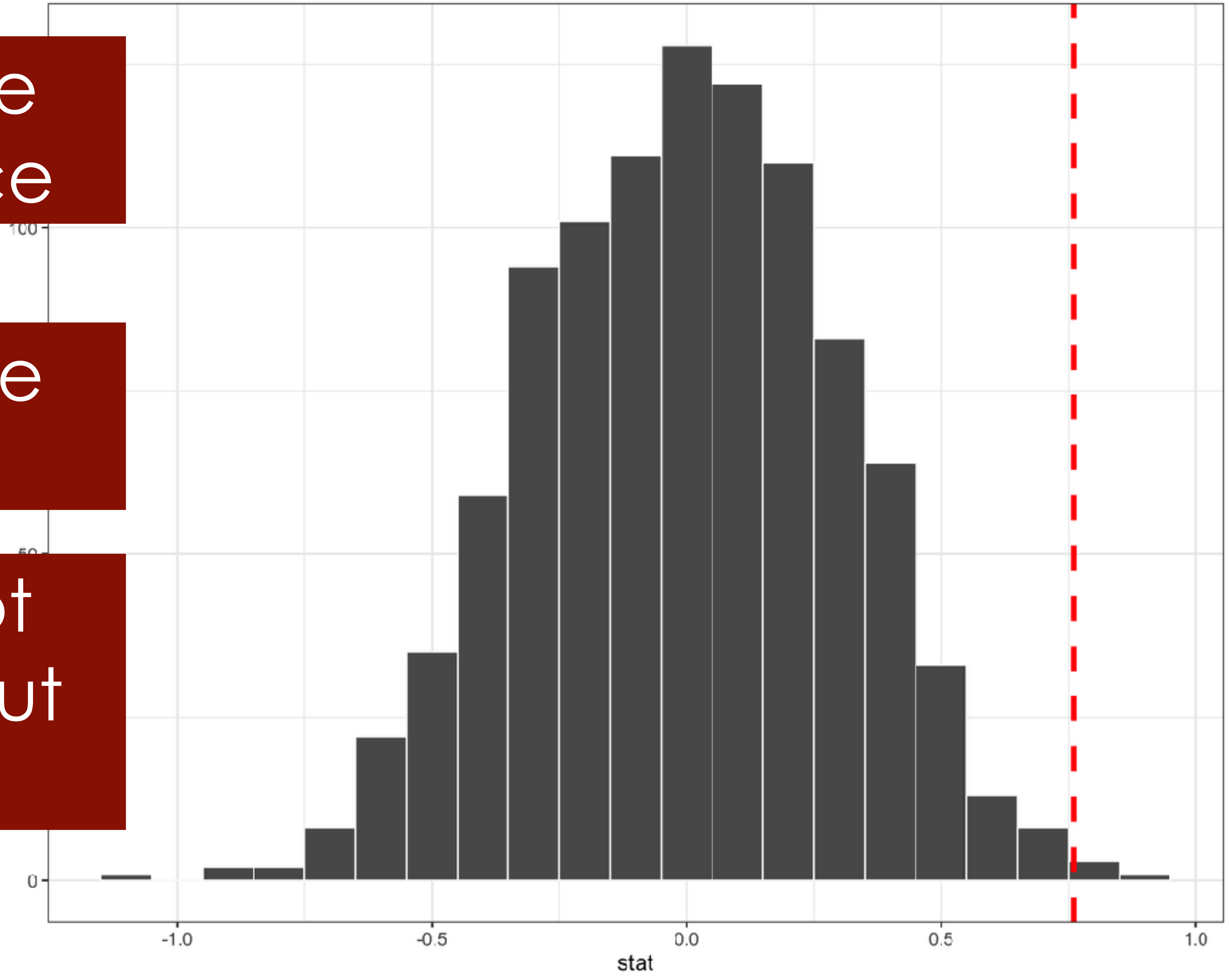
```
# permutation based hypothesis testing
null_affairs = Affairs %>%
  specify(formula = affairs ~ children) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("yes", "no"))
```

Permutation = by chance

Null distribution: what we might observe by chance

Sample statistic: what we actually observed

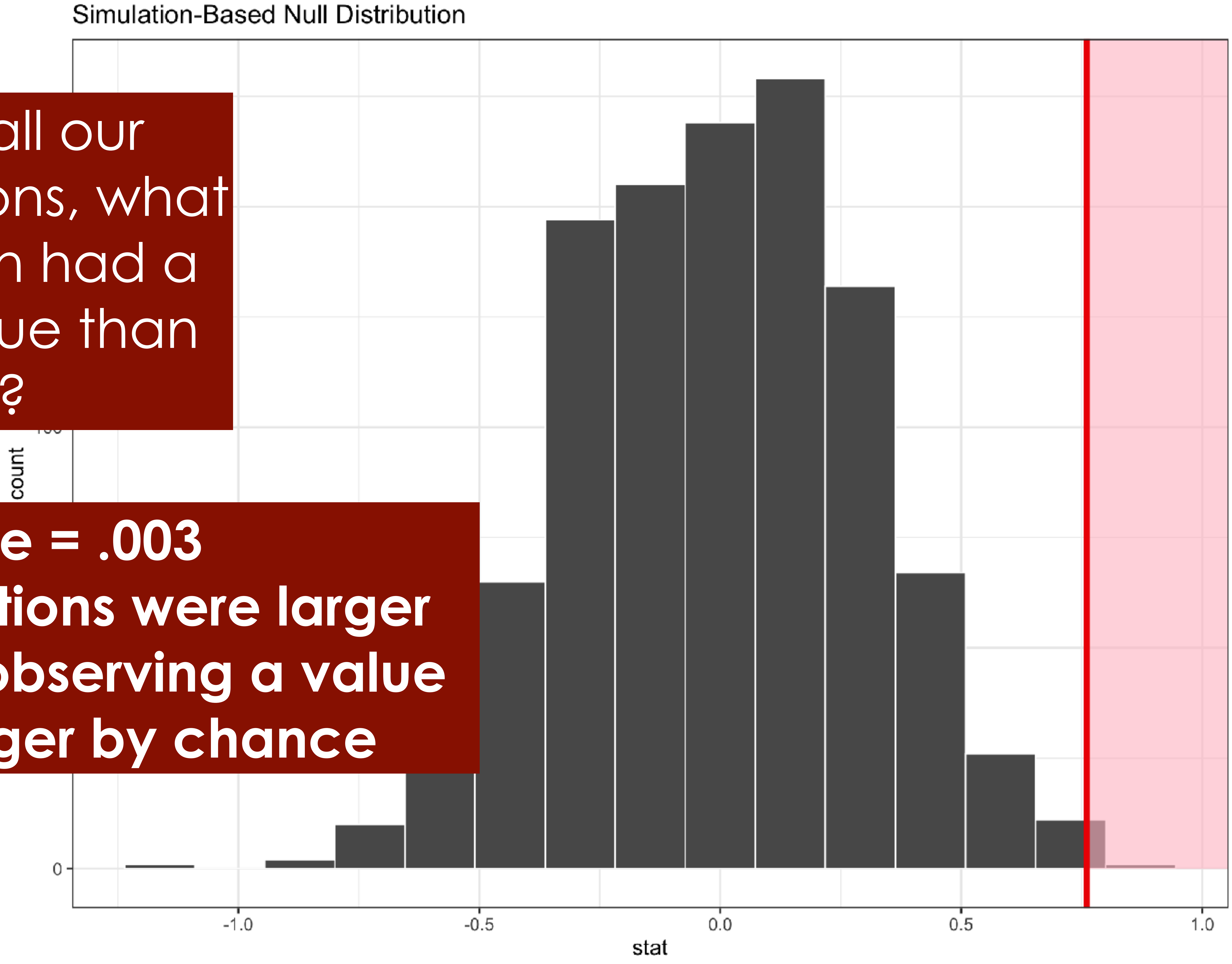
What we observed is not impossible by chance, but very unlikely



What do we mean by “very unlikely”?

P-values

The probability of observing an effect at least that large when no effect exists



Out of all our permutations, what proportion had a larger value than .76?

p-value = .003
.3% of permutations were larger
.3% chance of observing a value this big or larger by chance

OK, but how unlikely does something have to be before we reject the null ("declare guilty")?

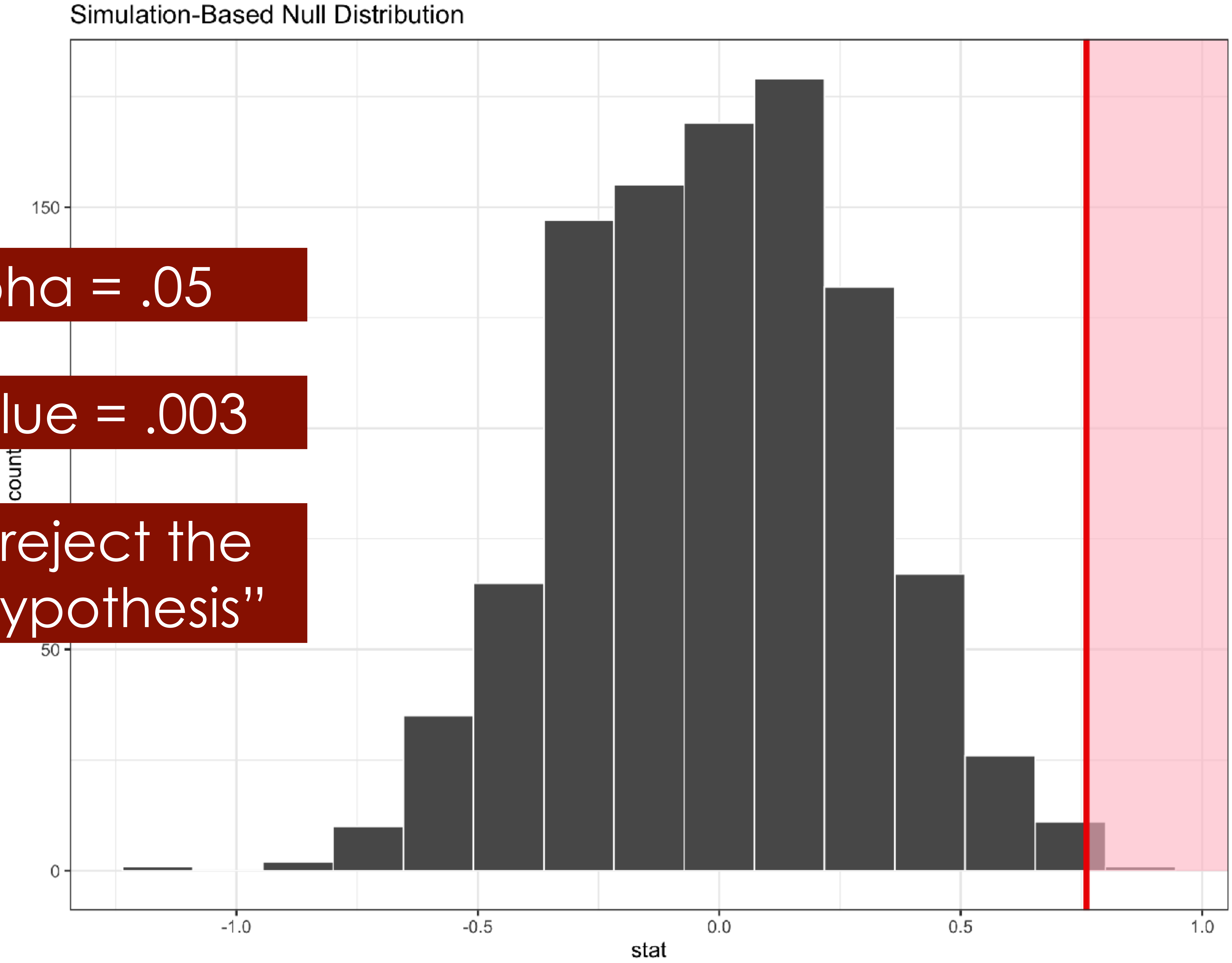
Statistical significance

A threshold for deciding if enough evidence to safely reject the null

How small should the p-value be before we safely reject the null?

This “how small” threshold is the **significance level**, denoted by **alpha**

Our hypothesis (H1)	Null hypothesis (H0)
P-value < alpha	*P-value > alpha*
“Reject the null”	*“Fail to reject the null”*
“guilty”	*“not guilty”*



Alpha = .05

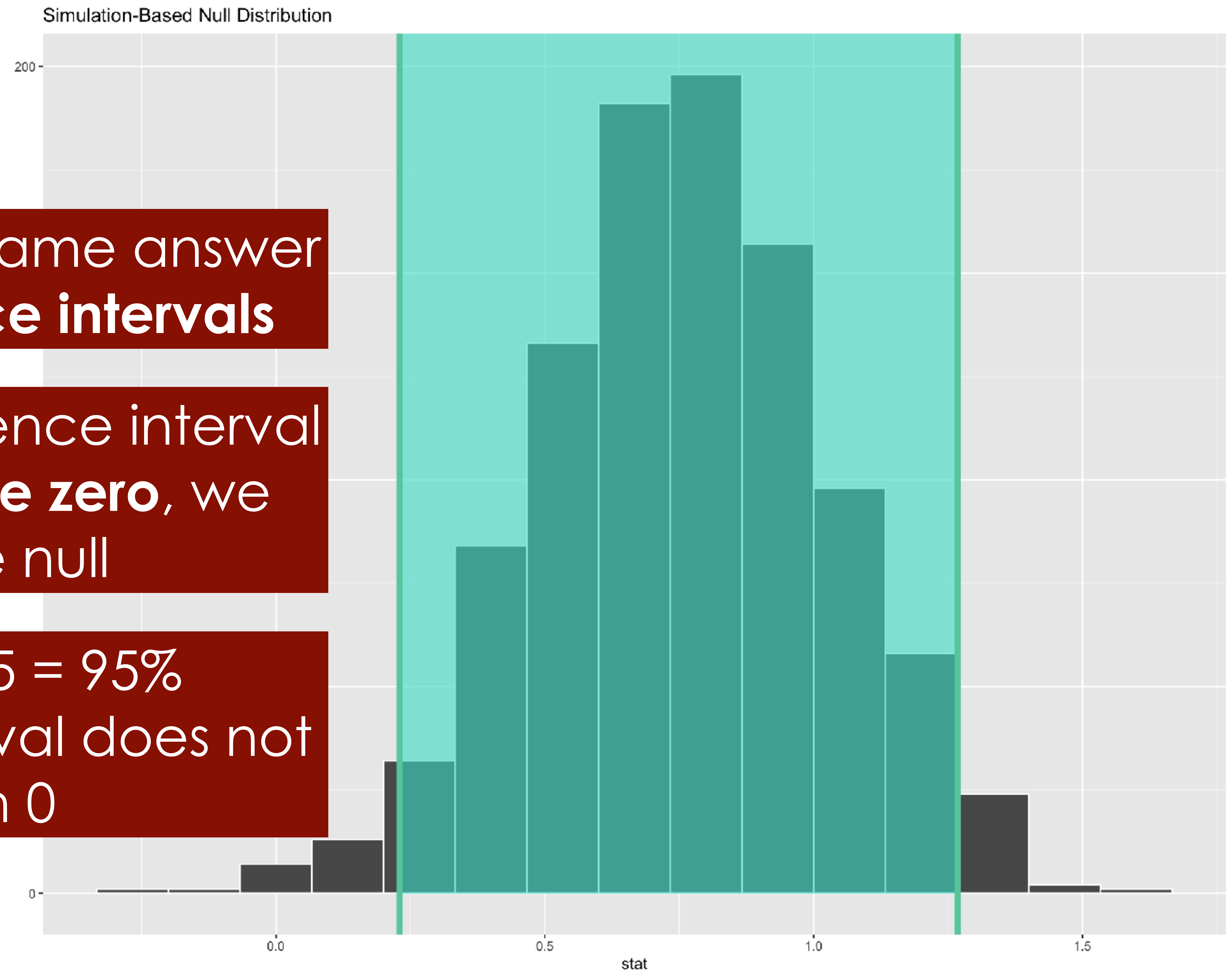
P-value = .003

"we reject the null hypothesis"

We can get the same answer using **confidence intervals**

If the 95% confidence interval **does not include zero**, we reject the null

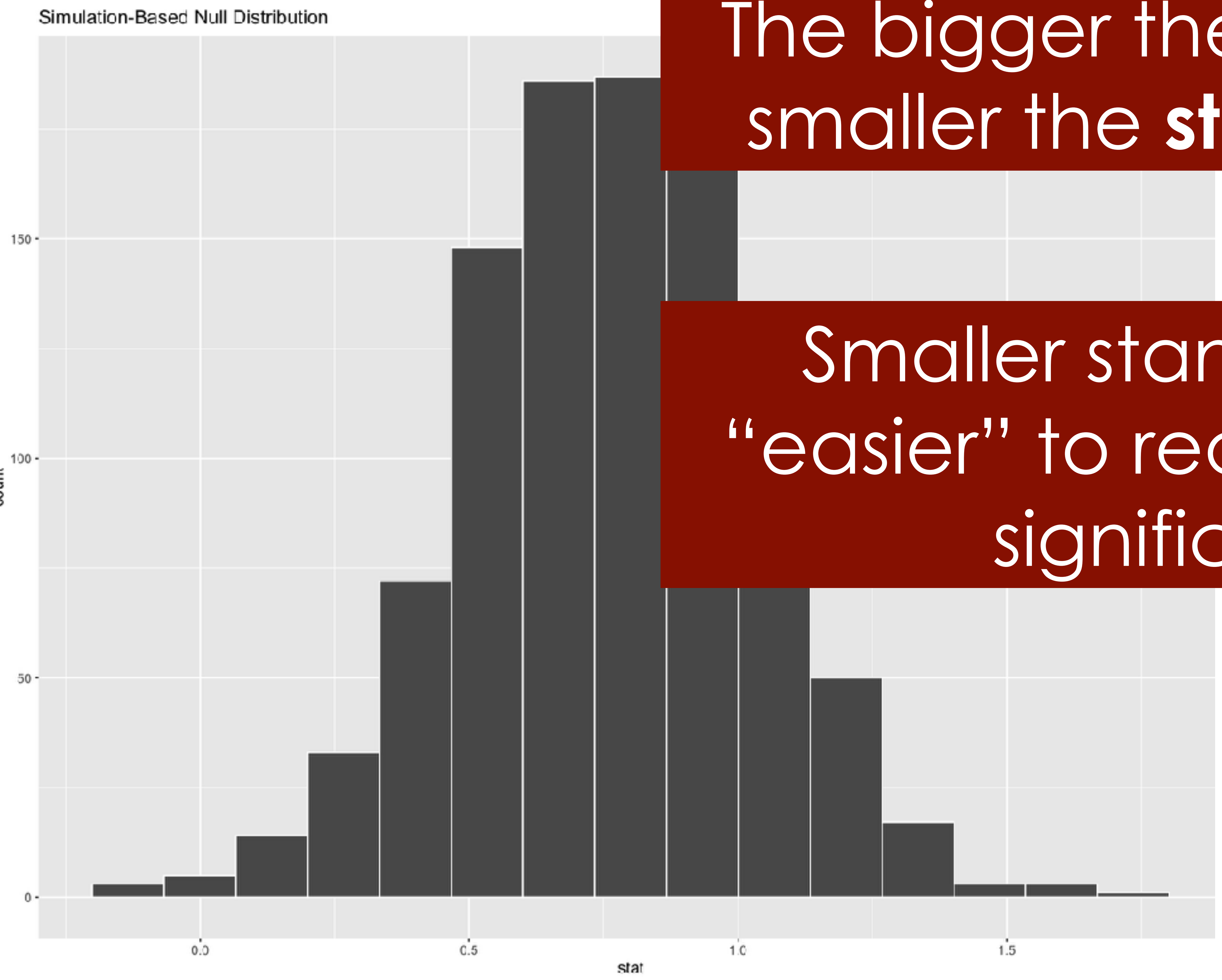
P-value < .05 = 95% confidence interval does not contain 0

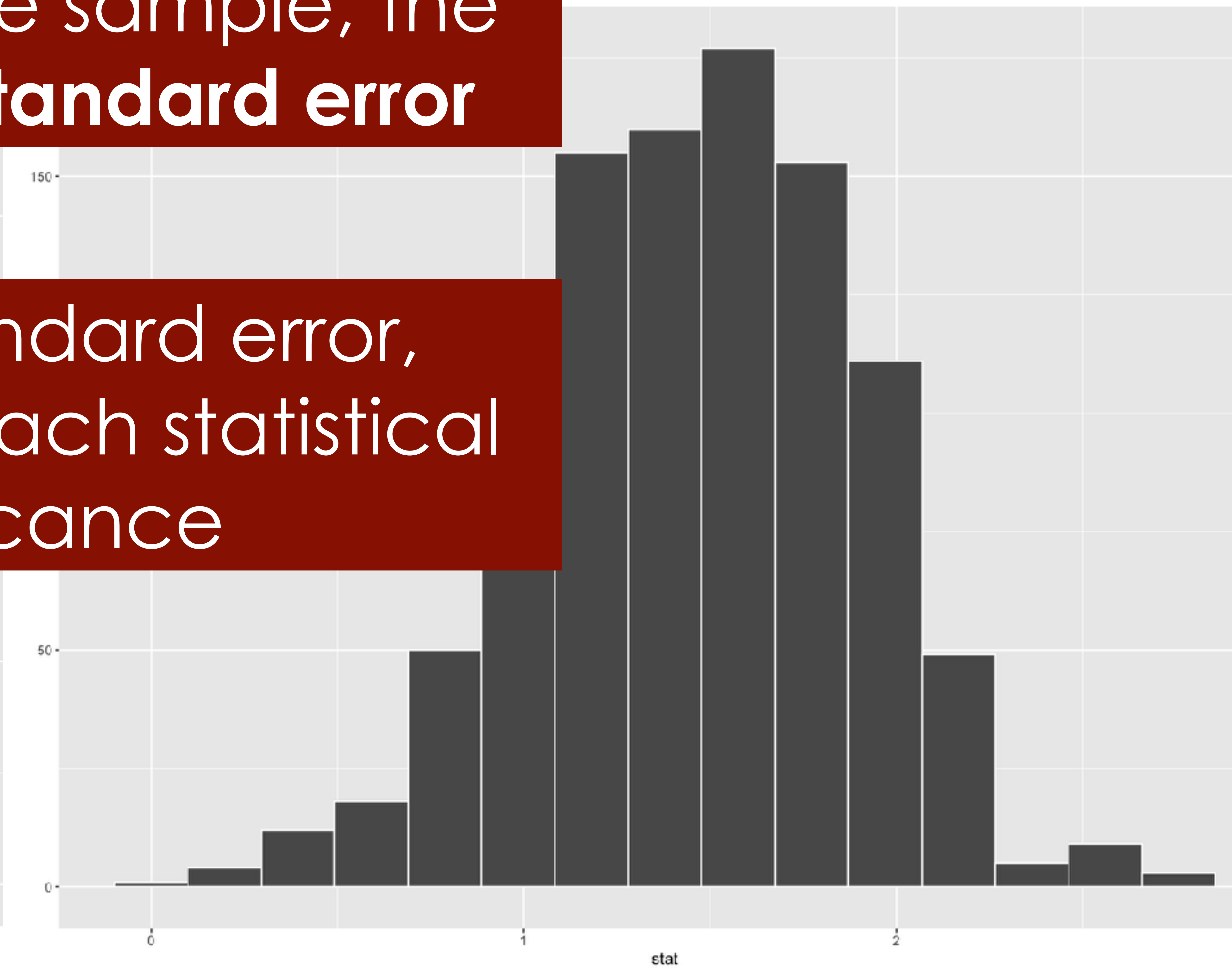


Sample size matters

The bigger the sample, the smaller the **standard error**

Smaller standard error, "easier" to reach statistical significance





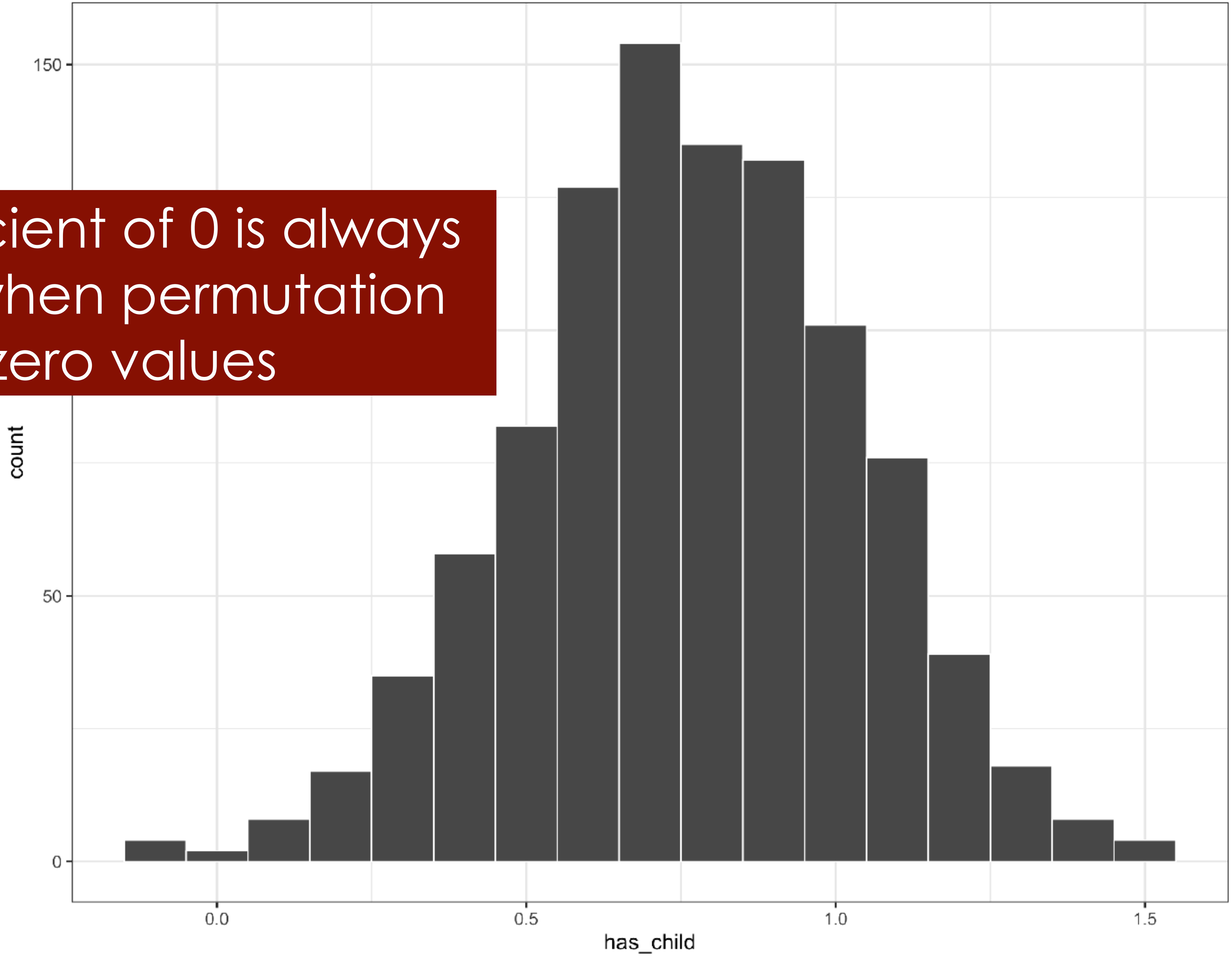
Choosing alpha

Alpha is typically .05

Why not set alpha = 0?

In p-value language: "0% chance of observing a value this large or larger by chance"

Getting a coefficient of 0 is always possible, even when permutation yields no zero values



The fact that 0 is always possible means we can't set alpha to 0

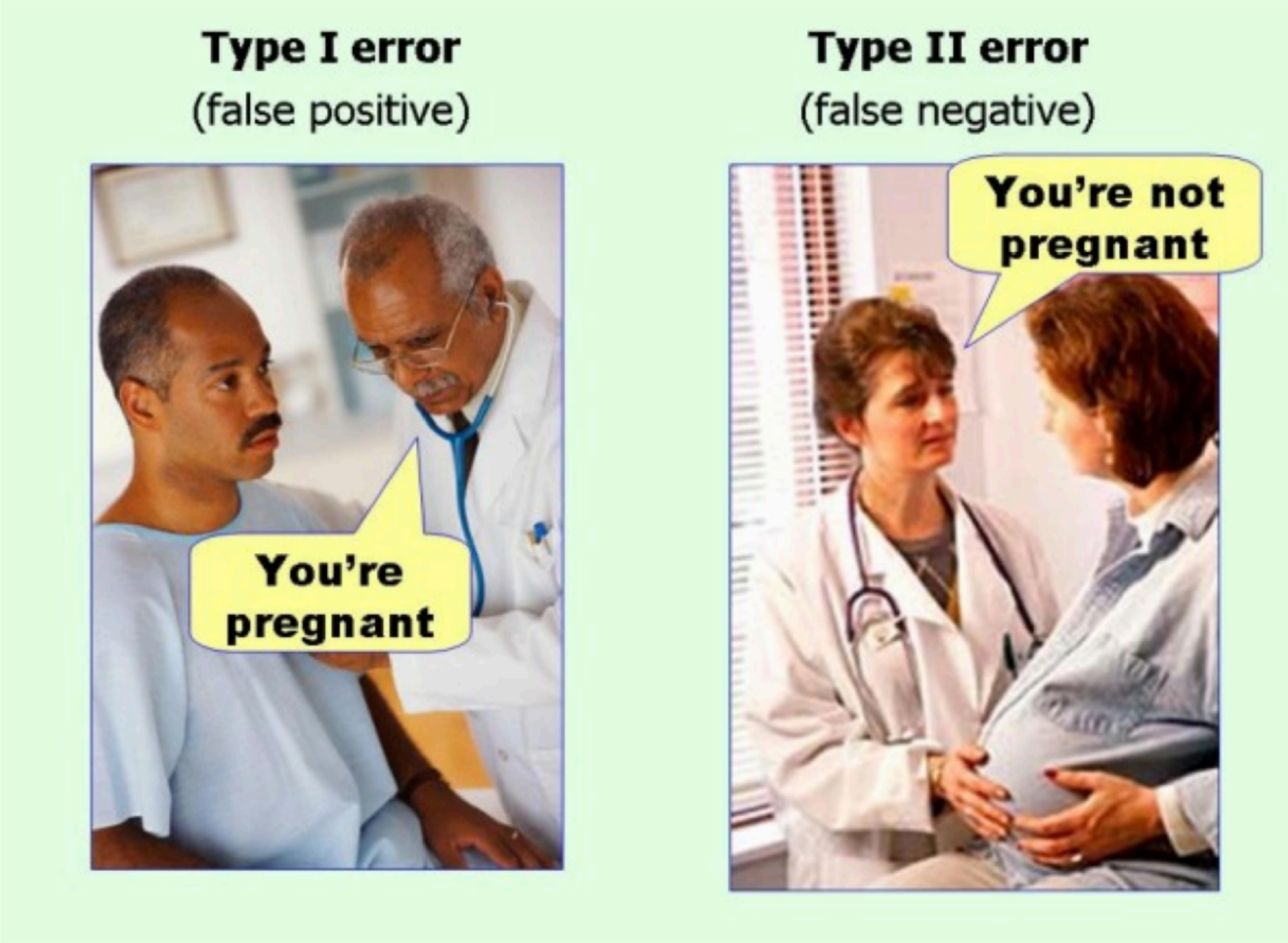
"No way to know with 100% certainty that null hypothesis is false"

"No way to know with 100% certainty that accused is guilty"

Why not then have a tiny alpha level, .00001, so we can be **really sure**?

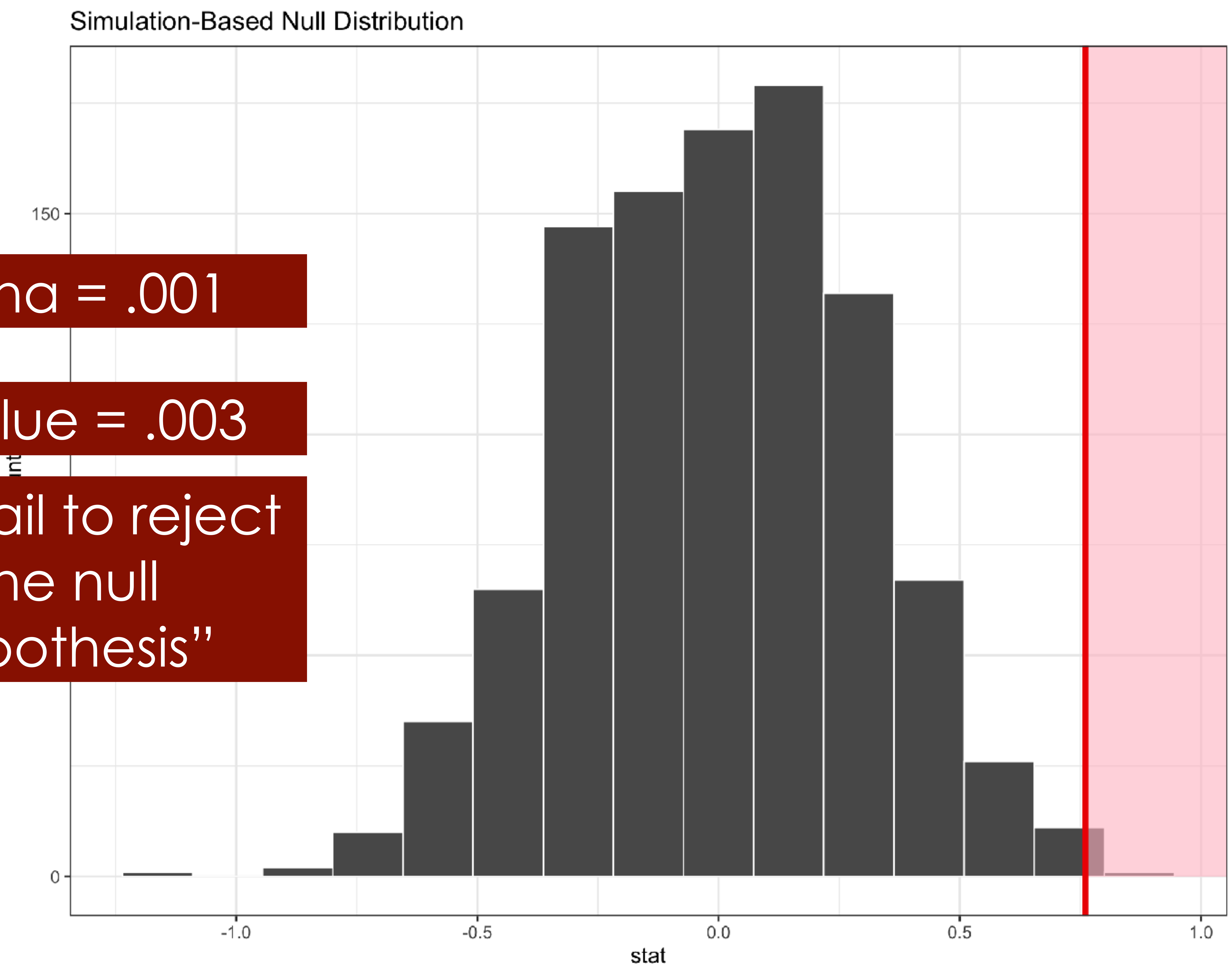
There's a tradeoff: small alpha = very unlikely we incorrectly reject null, but v**ery likely we incorrectly fail to reject null** (stats language is awful)

		Actual truth	
		Guilty	Not guilty
Jury decision	Guilty	Yay! True positive	Oh no! False positive (I)
	Not guilty	Oh no! False negative (II)	Yay! True negative


Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

		Actual truth	
		Yes effect	No effect
Result of hypothesis test	Yes effect	Yay! True positive	Oh no! False positive (I)
	No effect	Oh no! False negative (II)	Yay! True negative

α
0.10
0.05
0.01



Alpha = .001

P-value = .003

“we fail to reject the null hypothesis”

Choosing alpha levels

Alpha is our tolerance for rejecting H0 when in fact H0 is true (Type 1 Error)

The lower alpha is the less likely we are to make Type 1 error

But the more likely we are to make Type 2 error

Again, alpha typically = .05

Regression tables

```
# gapminder results
lm(gdpPercap ~ lifeExp + continent, data = gapminder) %>%
  get_regression_table()
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-16963	1059	-16.0	0	-19040	-14886
lifeExp	392	20.7	18.9	0	351	433
continentAmericas	-1249	642	-1.94	0.052	-2509	10.9
continentAsia	1318	556	2.37	0.018	227	2409
continentEurope	3244	706	4.59	0	1859	4629
continentOceania	6447	1720	3.75	0	3073	9821

Y = test scores

Not small vs. small

Class doesn't have aide vs. class has aide

Student not white/Asian vs. yes

Student is boy vs. girl

Student does not receive FRL vs. yes

Teacher not white/Asian vs. yes

Years (actual number)

Teacher does not have MA vs. yes

Table 2: OLS models for four standardized tests

VARIABLES	(1) Reading	(2) Math	(3) Listening	(4) Words
Small class	6.47***	8.84***	3.24**	6.99***
	(1.45)	(2.32)	(1.42)	(1.60)
Regular + aide class	1.00	0.42	-0.58	1.27
	(1.26)	(2.14)	(1.32)	(1.42)
White or Asian	7.85***	16.91***	17.98***	7.08***
	(1.61)	(2.40)	(1.70)	(1.91)
Girl	5.39***	6.46***	2.67***	5.03***
	(0.78)	(1.12)	(0.74)	(0.94)
Free/reduced lunch	-14.69***	-20.08***	-15.23***	-15.97***
	(0.91)	(1.33)	(0.90)	(1.07)
Teacher white or Asian	-0.56	-1.01	-3.68	0.46
	(2.66)	(3.80)	(2.59)	(3.07)
Years of teacher experience	0.30**	0.42**	0.25*	0.30**
	(0.12)	(0.20)	(0.15)	(0.14)
Teacher has MA	-0.75	-2.20	0.50	0.24
	(1.25)	(2.08)	(1.24)	(1.46)
School fixed effects	X	X	X	X
Constant	431.69***	475.52***	531.28***	428.97***
	(3.12)	(4.49)	(2.84)	(3.59)

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

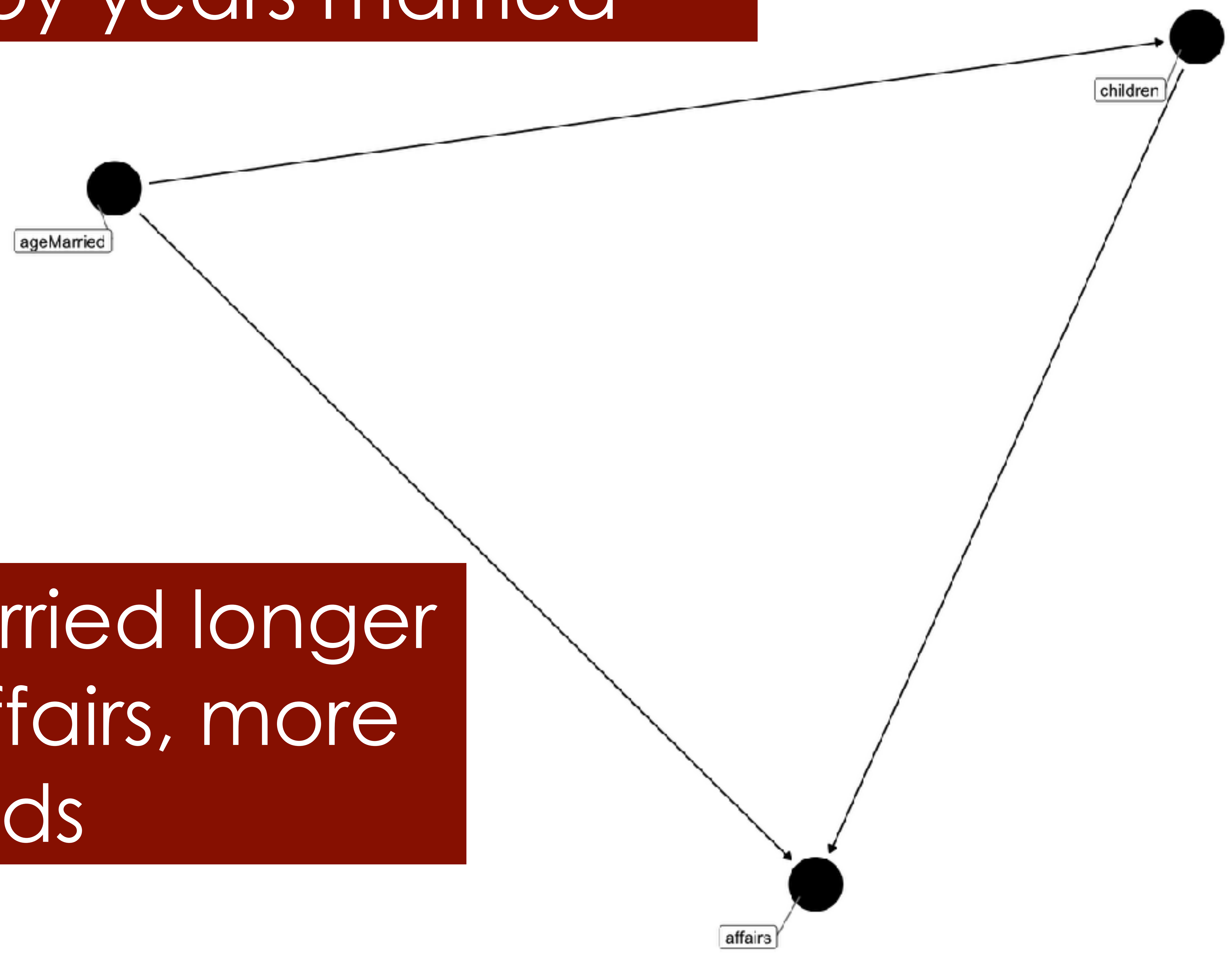
Remember the DAGs

Hypothesis testing is a useful way to deal with **the sample problem**

Gives us a standard for **quantifying certainty** in our estimates

But a statistically significant result does not mean our inferences are correct!

Effect of children on affairs likely confounded by years married



I.e., people who are married longer = more likely to have affairs, more likely to have kids

Notice what happens to coefficient on children

	(1)	(2)
(Intercept)	0.912 ***	0.562 *
	(0.251)	(0.264)
childrenyes	0.760 *	-0.033
	(0.297)	(0.358)
yearsmarried		0.112 ***
		(0.029)

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