Welcome to STA 101!

Midterm 2: November 12 - 14 (two weeks from today)

- **Same format**: 70% in-class (no tech); 30% take-home (tech);
  - In-class: mostly multiple choice;
  - In-class: fewer silly freebies;
  - In-class: visual intuition will be emphasized.

- **Same extra credit** (105 points available, graded out of 100):
  - $+1$ practice in-class
  - $+1$ practice take-home
  - $+1$ review sheet
  - $+1$ lab review attendance
  - $+1$ cheat sheet

- **Prepare**:
  - Complete all the extra credit assignments;
  - Carefully read Ch. 11, 12, 13, 14, (and 15, I guess);
  - Work odd-numbered exercises in the back of these chapters;

(if applicable, make appointments in the testing center now.)

What is an hypothesis test trying to do?

Simplest example: flipping an unfamiliar coin to determine if it's fair (equally likely to come up heads or tails).

Competing claims (hypotheses):

$$H_0 : p = 0.5 \qquad \text{(coin is fair)}$$
$$H_A : p \neq 0.5 \qquad \text{(coin is unfair)}$$

Result: we flip the coin a bunch of times and get 51% heads.

Fact: $51\% \neq 50\%$. So what?

Two possibilities:

- Fair coin. We just got 51% as a quirk of the random sampling;
- Unfair coin. $51\% \neq 50\%$. Anomaly detected! Case closed!

The whole ballgame: how do we tell the difference?

How do we tell the difference?

Two competing claims:

$$H_0 : p = p_0$$
$$H_A : p \neq p_0$$

Result: collect a random sample and compute estimate $\hat{p}$.

Problem: You probably won't get exactly $\hat{p} = p_0$. So what?

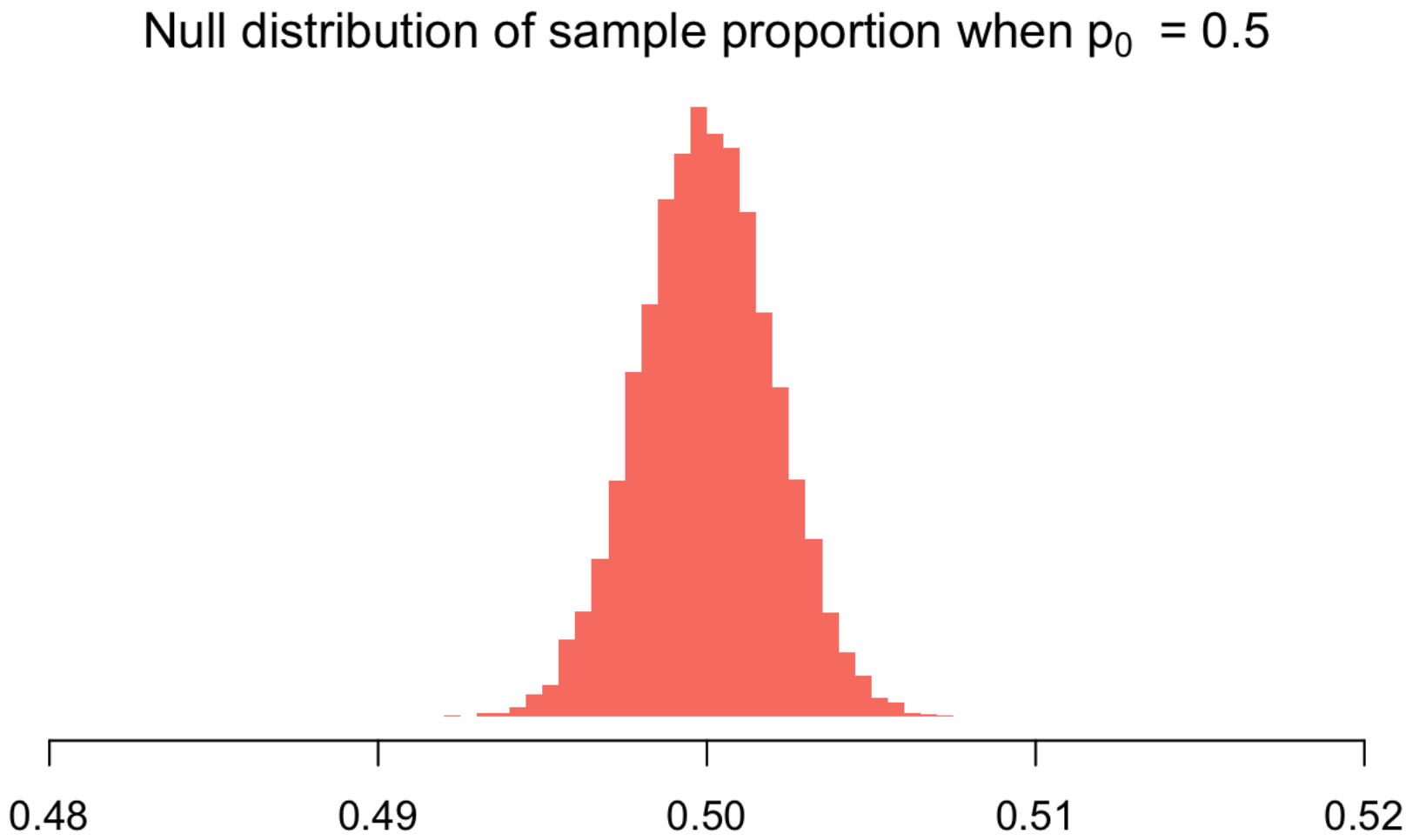
Two possibilities:

- $p = p_0$. We only got $\hat{p} \neq p_0$ because of random sampling;
- $p \neq p_0$. That's why $\hat{p} \neq p_0$. Take the hint, dummy.

In order to rule out the first possibility, we look at the variability in $\hat{p}$ that would be produced by random sampling *if the null were true*.

The null distribution

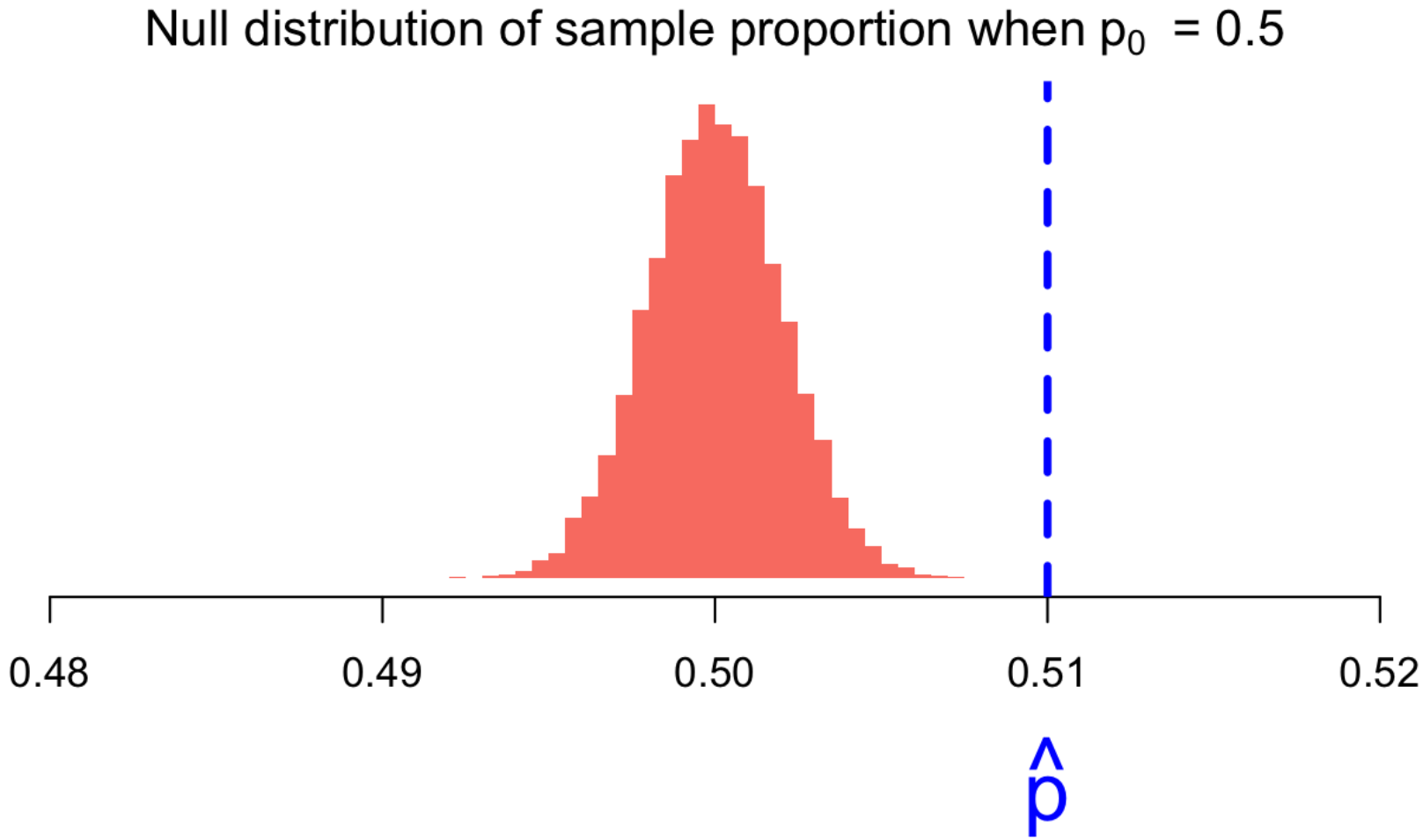
Assume the null hypothesis H_0 is true. In that case, what would the variation in the estimate $\hat{p}$ look like?



This is a *hypothetical* sampling distribution.

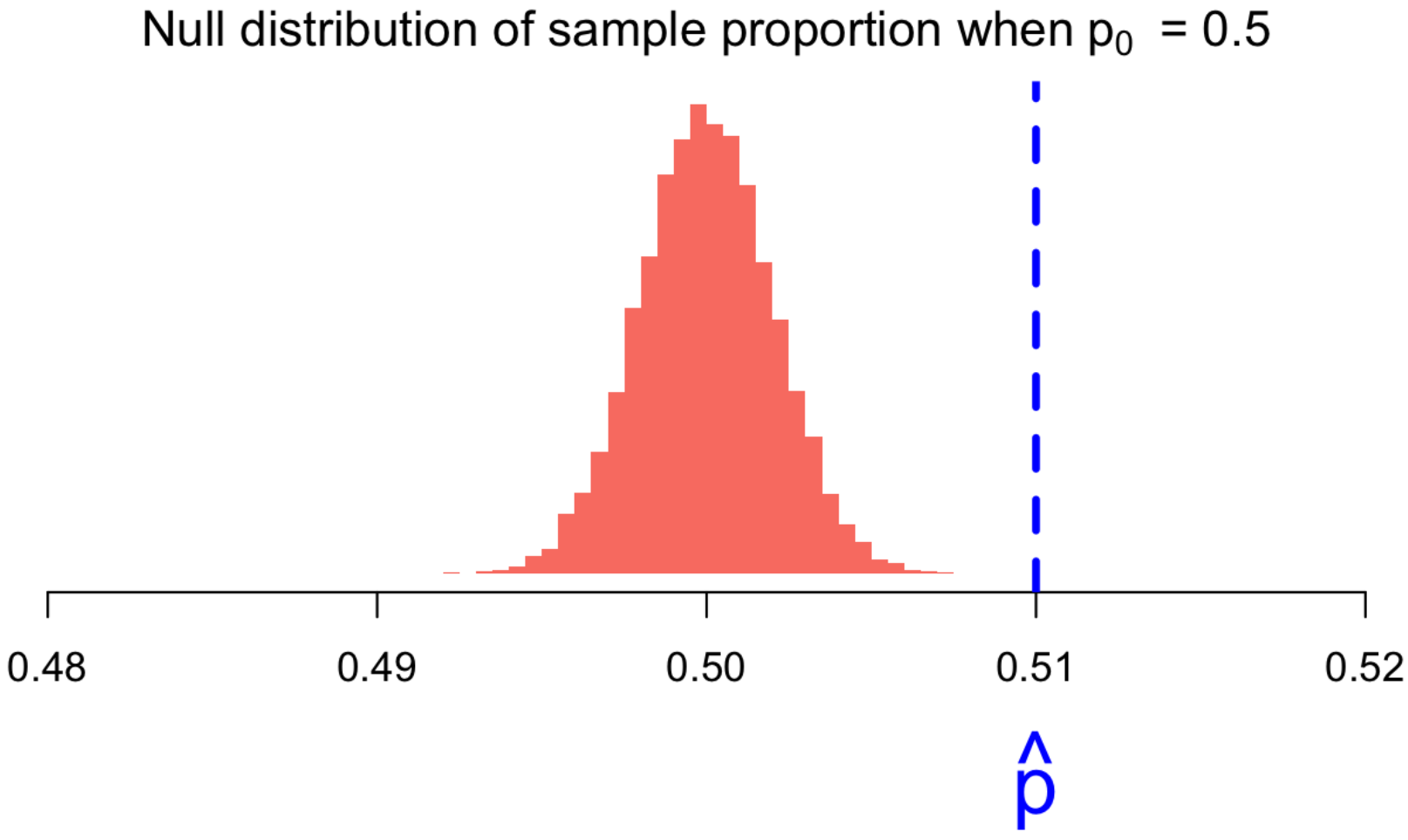
The null distribution

The null distribution is a hypothetical. Our actual data gave us an actual estimate. Find where it falls in the null distribution:

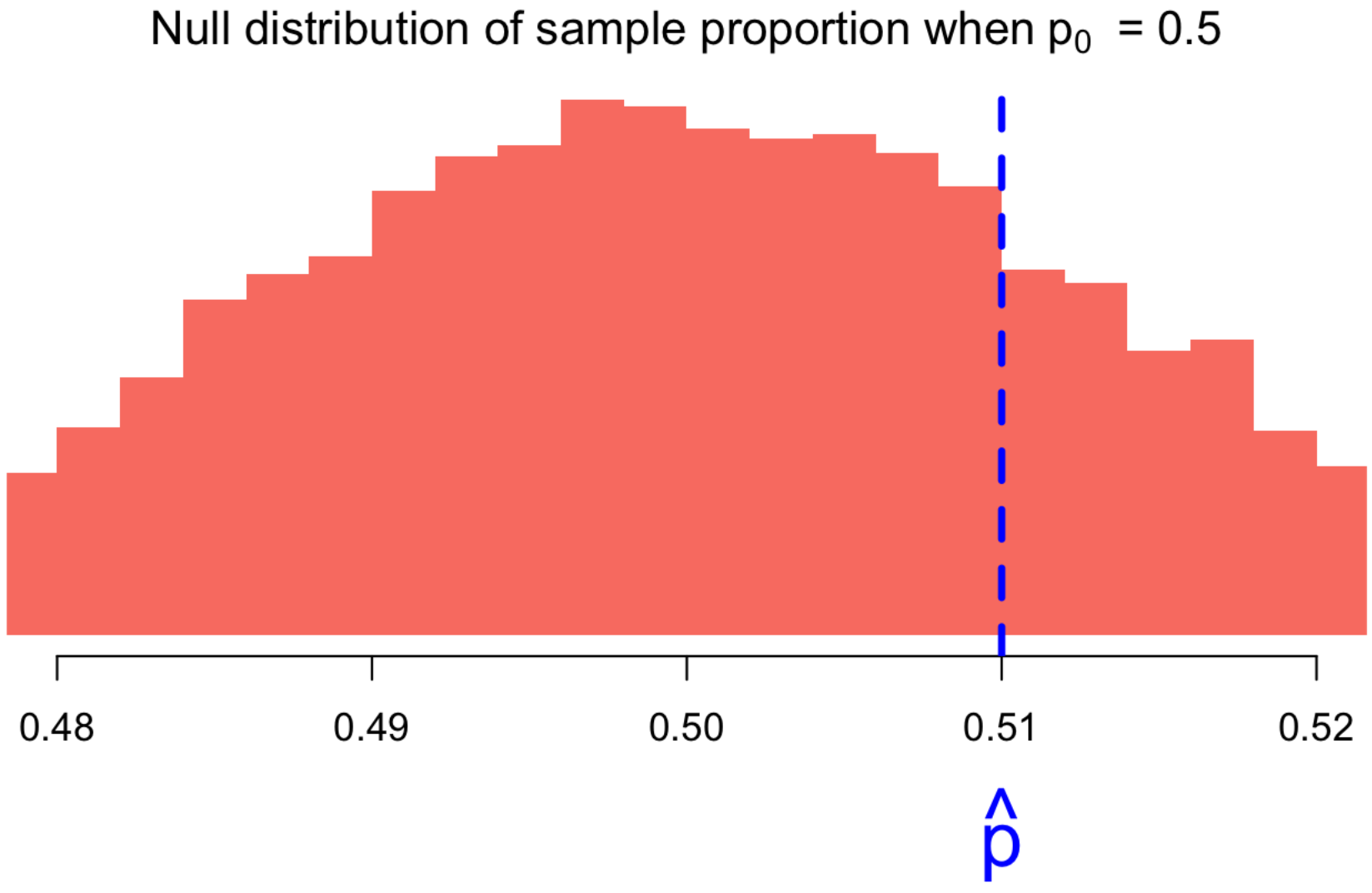


Question: Do reality (the actual estimate) and the hypothetical (the null distribution) look compatible, or not?

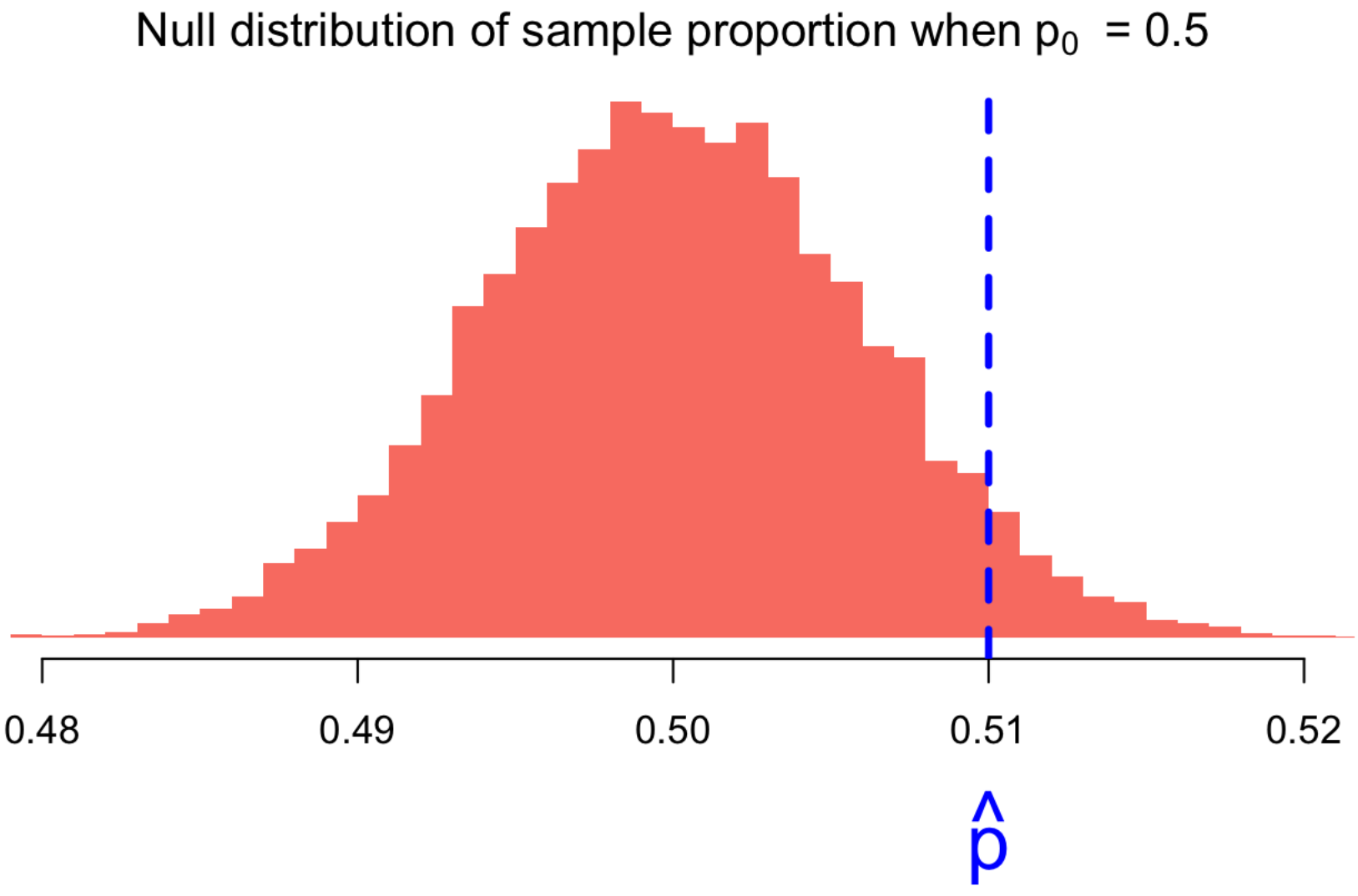
Probably incompatible (reject null)



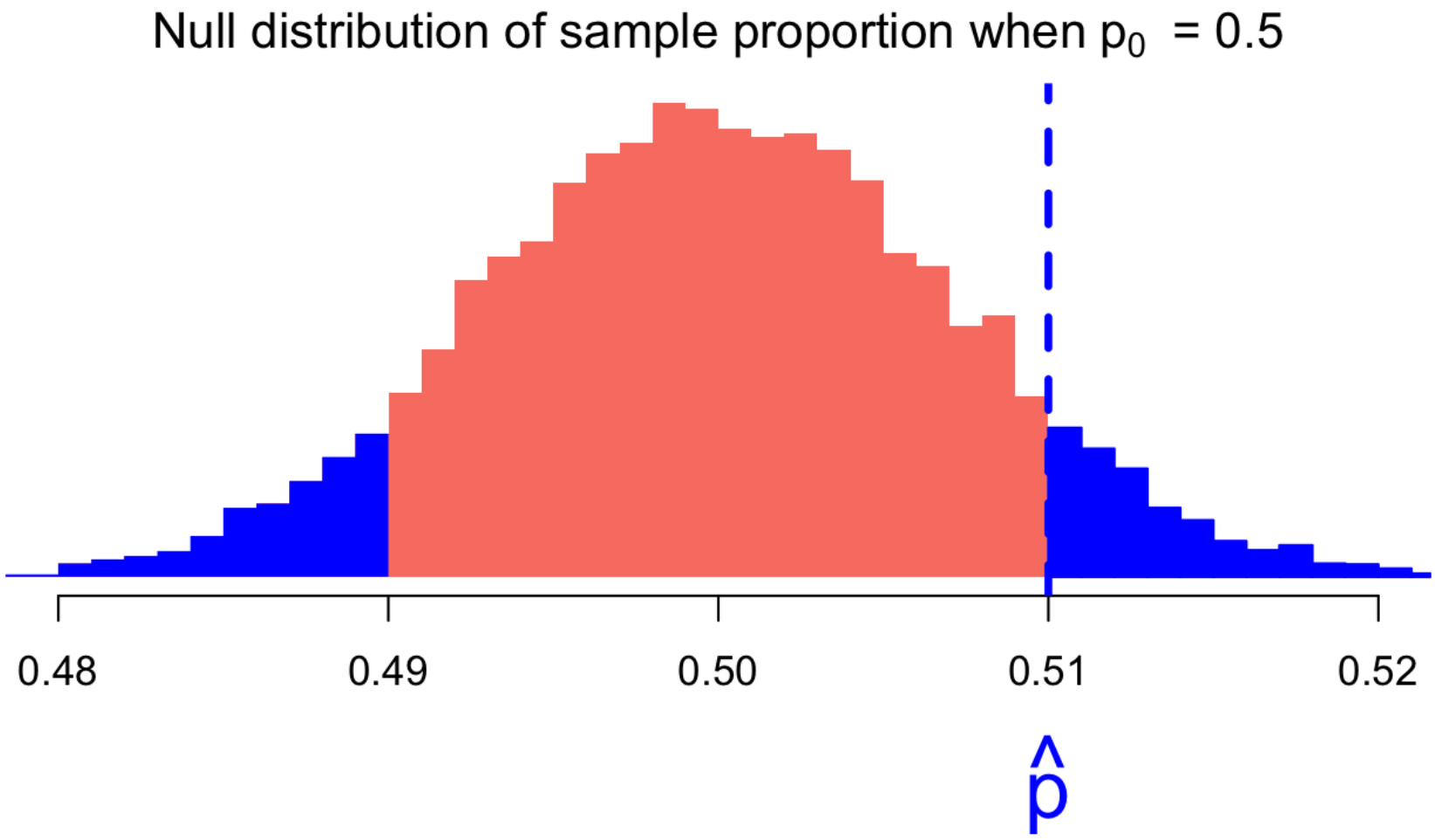
Probably compatible (fail to reject null)



Harder to tell



How do we quantify this comparison? *p*-value!

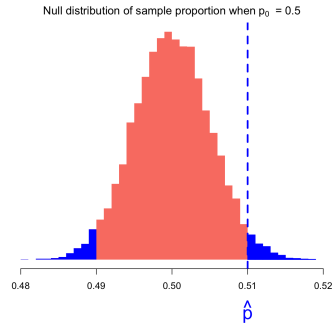


The *p*-**value** is the probability of an estimate as or more extreme than the one you actually got *if the null were true*. It's the proportion of the histogram area shaded blue.

How small is small?

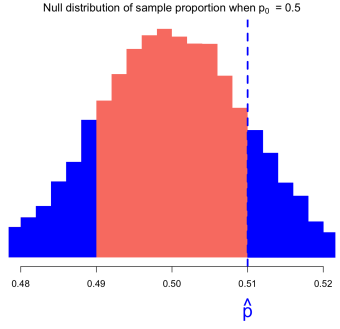
Smaller *p*-value:

- estimate is far in the tails of the null distribution;
- if H_0 true, your data/estimate would be nuts;
- reject the null.



Larger *p*-value:

- estimate is closer to middle of null distribution;
- if H_0 true, your data/estimate would be nbd;
- fail to reject the null.



Pick a cut-off/threshold $0 < \alpha < 1$:

- if *p*-value $< \alpha$, Reject H_0;
- if *p*-value $>= \alpha$, Fail to reject H_0.

α is called the **discernibility level**. How is it picked?

Recall: picking the confidence level of an interval estimate

Task: Choosing 75% vs 90% vs 95% vs 99% confidence?

Trade-off: We want an interval that is...

- ...wide enough to capture the truth with high confidence;
- ...narrow enough to teach us something meaningful about where the truth actually lives.

Silly example: The interval $(-\infty, \infty)$ is guaranteed to capture the truth 100% of the time. But it teaches us nothing.

Related: picking the discernibility level of a test

The choice of cut-off α can be domain and application dependent, but the overall goal is to balance the risk of two types of errors:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	Type 1 error	Correct!
	H_0 false	Correct!	Type 2 error

- Type 1 error = false positive;
- Type 2 error = false negative.

Example: a judge sentencing defendants

Hypotheses:

$$H_0 : \text{person is innocent}$$
$$H_A : \text{person is guilty}$$

Outcomes:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	1. Jail innocent person	Free innocent person
	H_0 false	Jail guilty person	2. Free guilty person

- Aspects of the American trial system regard a Type 1 error as worse than a Type 2 error (reasonable doubt standard, unanimous juries, presumption of innocence, etc).

Example: a doctor treating patients

Hypotheses:

$$H_0 : \text{person is well}$$
$$H_A : \text{person is ill}$$

Outcomes:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	1. Treat well person	Ignore well person
	H_0 false	Treat sick person	2. Ignore sick person

- Doctors tend to prefer treating too much than too little.

Example: the boy who cried wolf

Hypotheses:

$$H_0 : \text{no wolf}$$
$$H_A : \text{Run! A wolf!}$$

Outcomes:

		Your decision	
		Reject H_0	Fail to reject H_0
The truth	H_0 true	1. Panic over nothing	Go about your day
	H_0 false	Run from wolf	2. Get eaten

- In Part 1 of the story, townspeople commit a Type 1 error;
- In Part 2 of the story, townspeople commit a Type 2 error.

Picking the discernibility level of a test

Pick α to balance the risk of two types of errors:

		Your decision: Reject H_0	Your decision: Fail to reject H_0
The truth	H_0 true	Type 1 error	Correct!
	H_0 false	Correct!	Type 2 error

- $\alpha \uparrow \implies$ easier to reject H_0 $\implies$ Type 1 $\uparrow$ Type 2 $\downarrow$
- $\alpha \downarrow \implies$ harder to reject H_0 $\implies$ Type 1 $\downarrow$ Type 2 $\uparrow$
- **Typical choices**: $\alpha = 0.01, 0.05, 0.10, 0.15$.

One pathetic slide about power

Power is the probability of rejecting the null hypothesis when it is false (i.e. of avoiding a Type II error):

$$\text{Power} = \text{Prob}(\text{reject } H_0 \,|\, H_0 \text{ is false}).$$

It is the chance that a study will detect a deviation from the null if one really exists. We want this to be as big as possible.

Power is a function of

- Sample size;
- Deviation from the null one hopes to detect;
- Variability in your data;
- The discernibility level you choose.

Big ol' question: subject to constraints like budget, how should I design my study, and how much data should I collect, to make power as big as possible? *Very important, but beyond our course...*

Cardinal Sins in Statistics, Part 2 of 91

Thou shalt not interpret the p-value as the probability that the null hypothesis is true. It is the probability of an extreme result *assuming the null is true*.

Cardinal Sins in Statistics, Part 3 of 107

Thou shalt not confuse statistical discernibility with substantive importance.

What we say in STA 101	What you will hear elsewhere
discernibility level	"significance" level
statistical discernibility	statistical "significance"

Traditionally, if p-value $< \alpha$, we reject H_0 and call the result "statistically significant". But this wording often misleads people into thinking the results are just plain *significant*, in a substantive sense. **BOO! ICK! FALSE! WRONG! GO HOME!**

Example

The truth: a coin flip comes up heads with probability 0.499.

Hypotheses:

$$H_0 : \text{Prob(heads)} = 0.5$$
$$H_A : \text{Prob(heads)} \neq 0.5$$

Fact: H_0 is literally false. $0.5 \neq 0.499$.

Result: you flip the coin 10,000,000 times and get a *p*-value that's practically zero, and *correctly* reject H_0. So what?

Punchline: the machinery of statistics *cannot* tell you if your results are "meaningful" and "important". It can only tell you if the results are likely or not under random sampling.

statistical "significance" $\neq$ importance

Cardinal Sins in Statistics, Part 4 of 284

Thou shalt not *accept* the null hypothesis, even if the p-value is huge. You only "fail to reject" the null hypothesis.

Example: when a verdict is read out in court, it isn't "guilty" or "innocent." It's "guilty" or "not guilty," which is very different.

Hypothesis testing: an avalanche of itchy jargon

- **null hypothesis**
- **alternative hypothesis**
- **null distribution**
- ***p*-value**
- **discernibility level**
- **Type 1 error**
- **Type 2 error**
- **Power**
- ...and more.

Get it all on your cheat sheet!

Notice a pattern?

Statistical questions...

- **Q**: Does the linear model fit well?

  A: Look at the spread of the residual distribution.

- **Q**: Is the unknown parameter reliably estimated?

  A: Look at the spread of the sampling distribution.

- **Q**: Do we have sufficient evidence to reject the null?

  A: Look at the spread of the null distribution.

We typically represent a distribution with a histogram, and we measure spread with variance or standard deviation.

Make sure you understand these things! It's Chapter 5.

Into the weeds: how is the null distribution simulated?