

YData: Introduction to Data Science



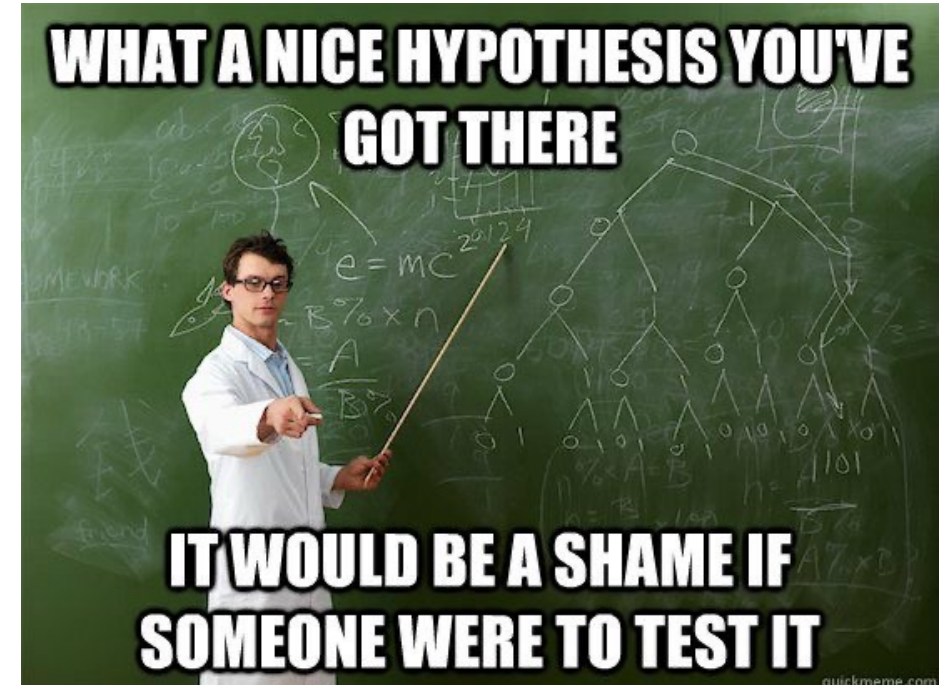
Class 22: Confidence intervals

Overview

Hypothesis tests for two means continued

Confidence intervals

If there is time: introduction to machine learning



Project timeline

~~Sunday, April 7th~~

- ~~• Projects are due on Gradescope at 11pm~~
- Also, email a pdf of your project to your peer reviewers
 - A list of whose paper you will review is on Canvas
 - Fill out the draft reflection on Canvas

Wednesday, April 17th

- Jupyter notebook files with your reviews need to be sent to the authors
- A template for doing your review is available

Sunday, April 28th

- Project is due on Gradescope
 - Add peer reviews to the Appendix of your project



Project peer review

A template for your project peer review has been posted

- `import YData`
- `YData.download.download_class_file('reviewer_template.ipynb', 'homework')`

Please review the projects by 11pm on Wednesday April 17th and:

- 1. Post a **pdf** of each of your reviews to Gradescope
- 2. Send a filled out **Jupyter Notebook** with your review to the project author
 - If you run into any logistic issues post to Ed and then ask our course manager Ashley (ashley.oaks@yale.edu)


In your final project, please add the three reviews in the Appendix section, and discuss how you addressed the reviewers' comments.

Also, homework 8 is due on Sunday April 14th

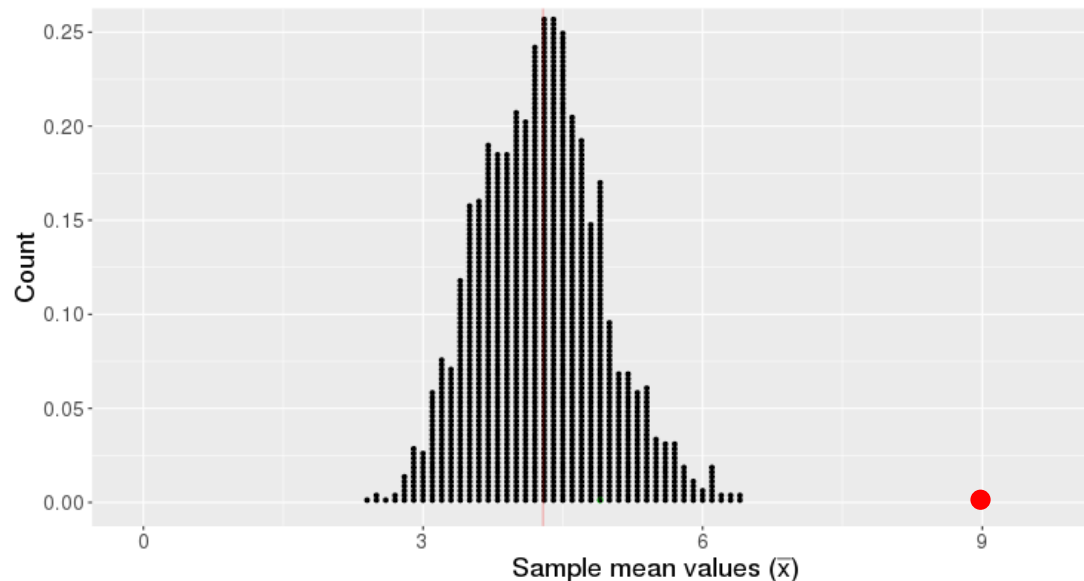
Hypothesis tests

Basic hypothesis test logic

We start with a claim about a population parameter

- E.g., $\mu = 4$ 

This claim implies we should get a certain distribution of statistics



If our observed statistic is highly unlikely, we reject the claim

Null and Alternative hypotheses

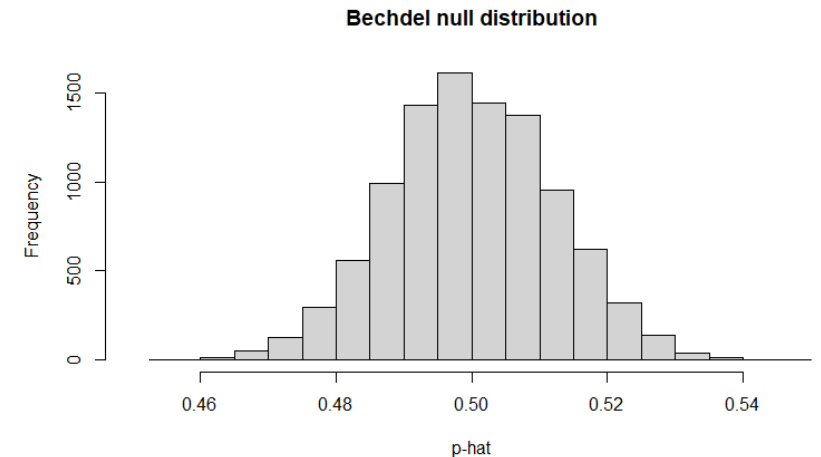
Null hypothesis

- A hypothesis where “nothing interesting” happened
 - E.g., our experiment failed
 - E.g., $H_0: \pi = 0.5$
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics

Alternative hypothesis

- The hypothesis we believe in (would like to see true)
- E.g., $H_A: \pi < 0.5$

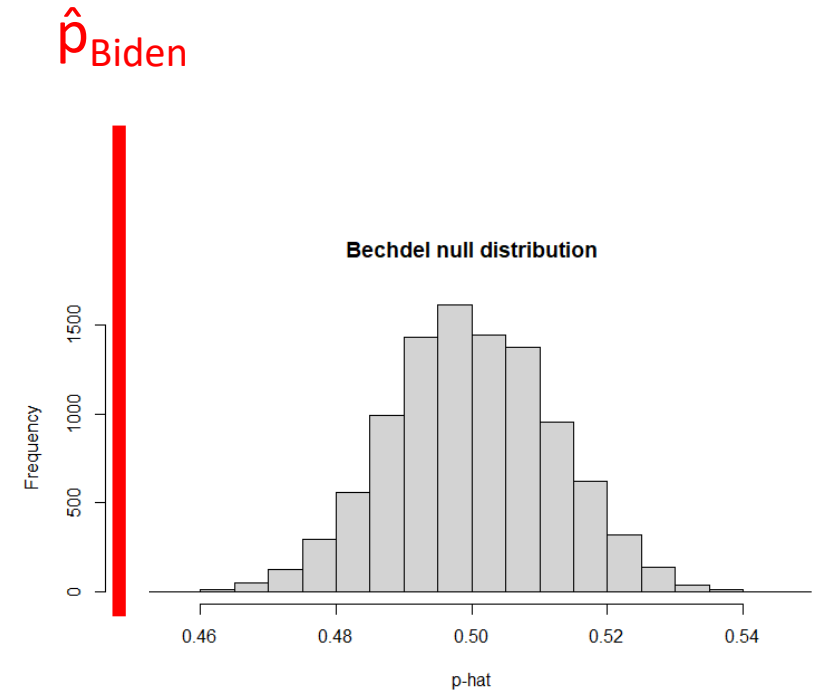
A **test statistic** is the statistic we choose to simulate in order to decide between the two hypotheses



Testing the null hypothesis

To resolve choice between null and alternative hypotheses:

- We compare the **observed test statistic** to the statistic values in the null distribution
- If the observed statistic is not consistent with the null distribution, then we can **reject the null hypothesis**
 - E.g., $H_0: \pi = 0.5$
- And we accept the alternative hypothesis
 - E.g., $H_A: \pi < 0.5$



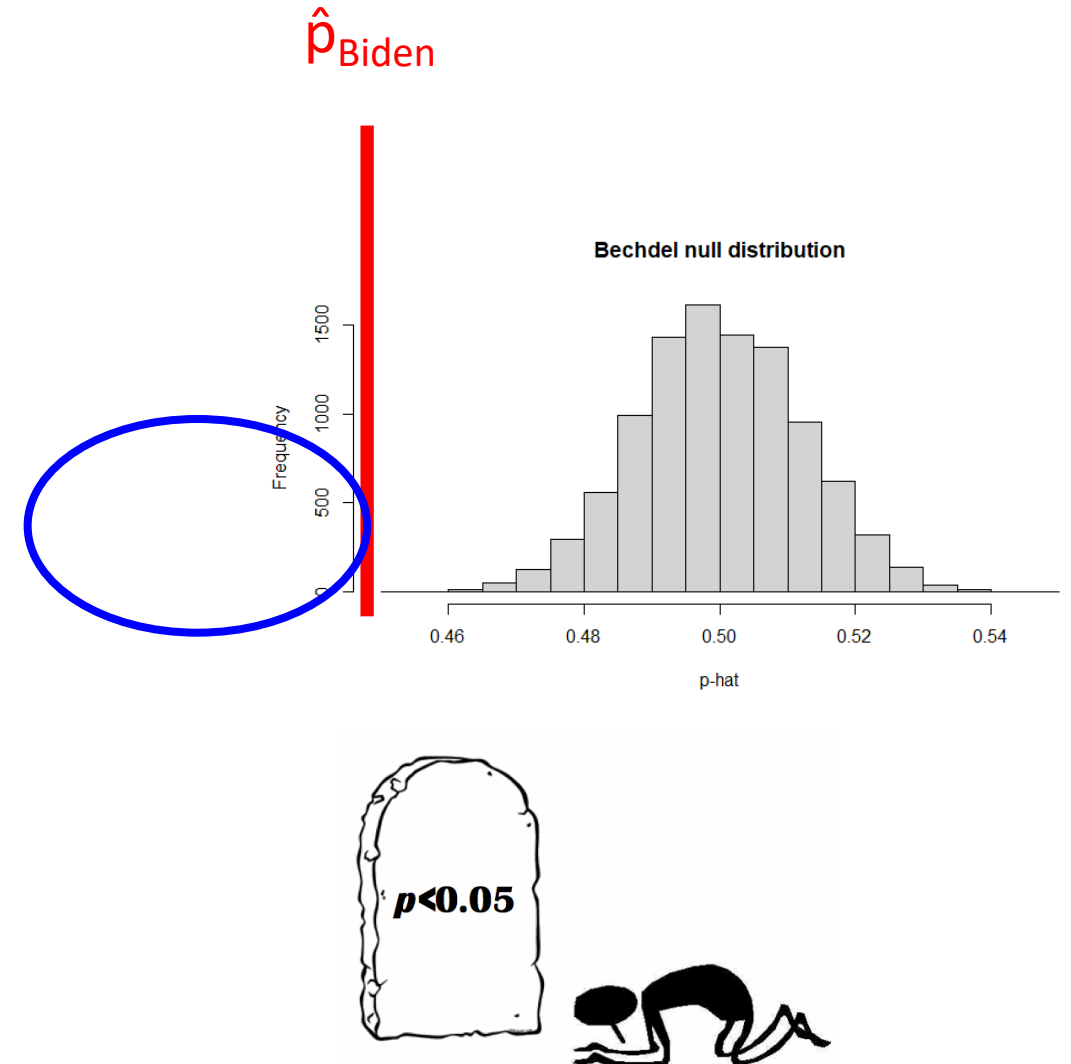
The p-value

The **p-value** is the probability, that we get a statistic as or more extreme than the observed statistic from the null distribution

- $P(\text{Null_Stat} \leq \text{obs_stat} \mid H_0)$

If the P-value is small, this is evidence against the null hypothesis and the results are often called "statistically significant"

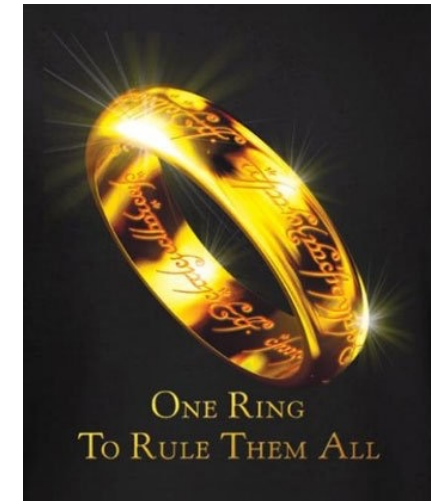
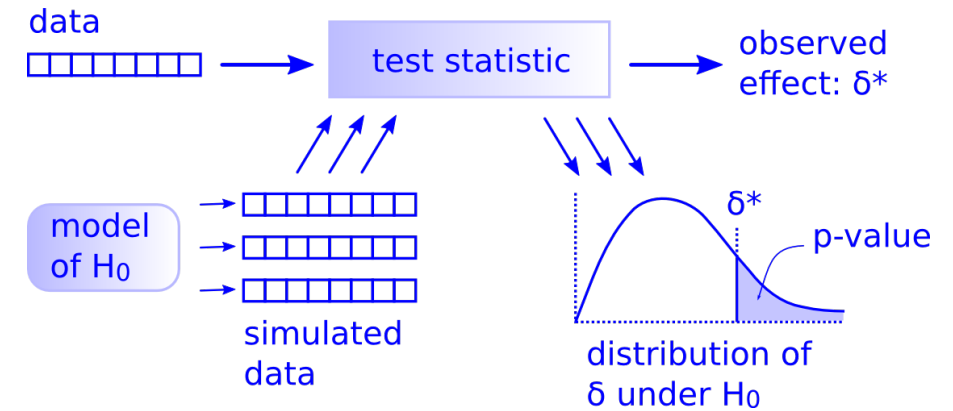
- Convention, $p\text{-value} < 0.05$



Steps needed to run a hypothesis test

To run a hypothesis test, we can use 5 steps:

1. State the null and alternative hypothesis
2. Calculate the observed statistic of interest
3. Create the null distribution
4. Calculate the p-value
5. Make a decision



Baby birth weights

Question: Is the average weight of babies at birth affected by whether a mother smokes?

To gain insight into this question let's compare:

- A. Birth weights of babies of mothers who smoked during pregnancy
- B. Birth weights of babies of mothers who didn't smoke



Step 1: State the null and alternative hypotheses

Null hypothesis:

- In the population, the average birth weights of the babies in the two groups are the same.

Alternative hypothesis:

- In the population, the babies of the mothers who didn't smoke were heavier, on average, than the babies of the smokers.



How can we write these hypotheses using symbols we have discussed?

$$H_0: \mu_{\text{non-smoke}} = \mu_{\text{smoke}}$$

or

$$\mu_{\text{non-smoke}} - \mu_{\text{smoke}} = 0$$

$$H_A: \mu_{\text{non-smoke}} > \mu_{\text{smoke}}$$

or

$$\mu_{\text{non-smoke}} - \mu_{\text{smoke}} > 0$$

Step 2: Compute the observed statistic

Let's look at a data set from 1236 mother-baby pairs that was collected between 1960 and 1967 among women in the Kaiser Foundation Health Plan in the San Francisco East Bay area

- 742 mothers who did not smoke
- 484 mothers who smoked

Statistic: Difference between average birth weights

- $\bar{x}_{\text{non-smokers}} - \bar{x}_{\text{smoker}}$

Large values of this statistic favor the alternative



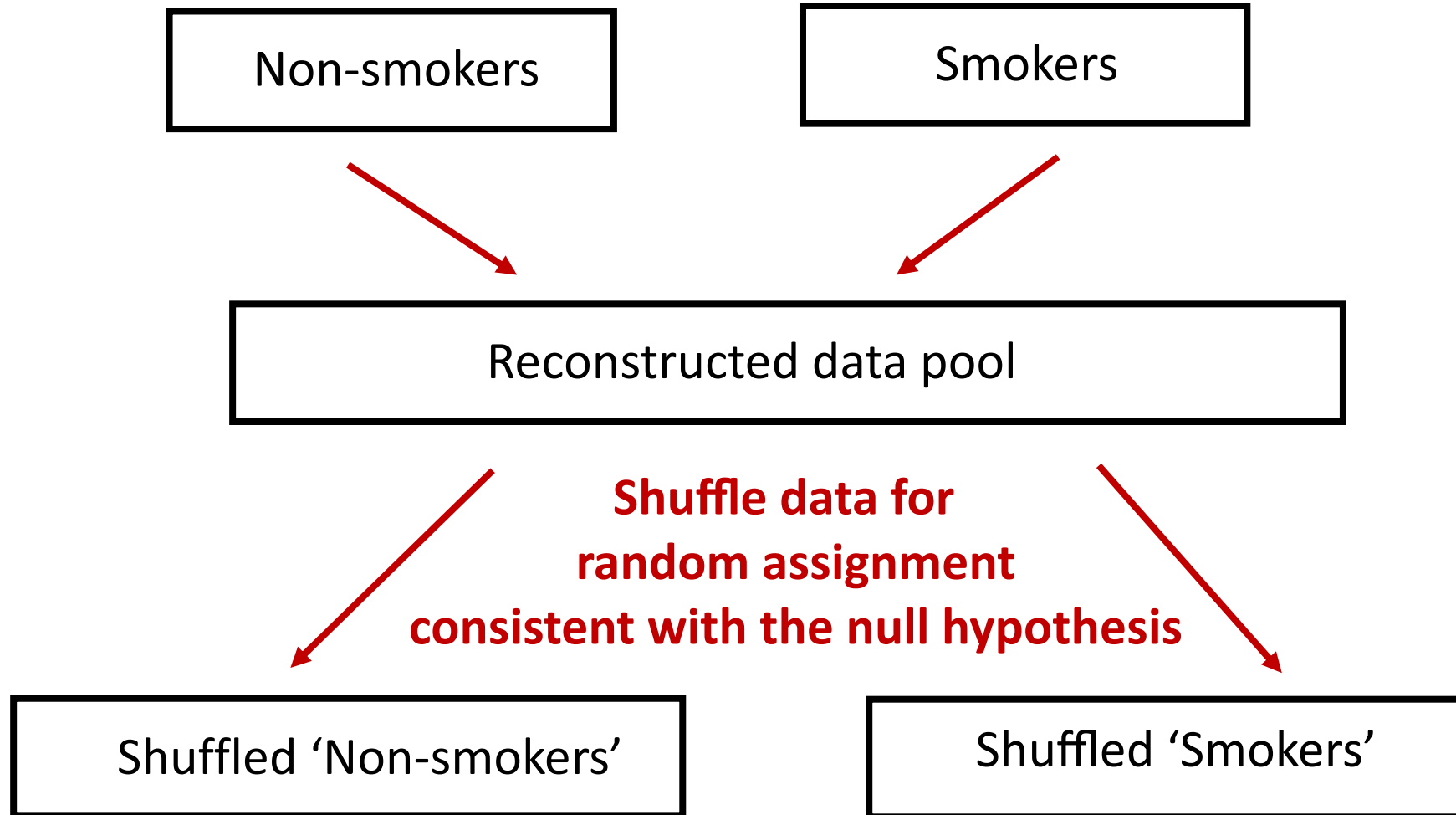
Step 3: Create the null distribution

If the null is true, all rearrangements of the birth weights among the two groups are equally likely

Plan:

- Shuffle all the birth weights
- Assign some to "shuffled smokers" and the rest to "shuffled non-smokers", maintaining the two sample sizes
- Find the difference between the averages of the two shuffled groups
- Repeat

Create the null distribution!



One null distribution statistic: $\bar{x}_{\text{shffle-non-smokers}} - \bar{x}_{\text{shuffle-smoker}}$



Let's explore this in Jupyter!

Brief mention: two-sided hypothesis tests

So far we have always had a specific prediction for the effect we observed

For example:

- We believed that *less than* 50% of movies passed the Bechdel test
- We believed that babies or mothers who did not smoke would way *more* (on average) than babies of mothers who smoked

This directionality was reflected in our alternative hypotheses

- $H_A: \pi_{\text{Bechdel}} < .5$
- $H_A: \mu_{\text{non-smoke}} > \mu_{\text{smoke}}$

Brief mention: two-sided hypothesis tests

Sometimes we do not know the direction of an effect, we only know that the value specified in the null hypothesis is not correct

For example:

- We just know that 50% of movies do not pass the Bechdel test
 - But it could be than more 50% or less than 50%

We would then write our alternative hypotheses as:

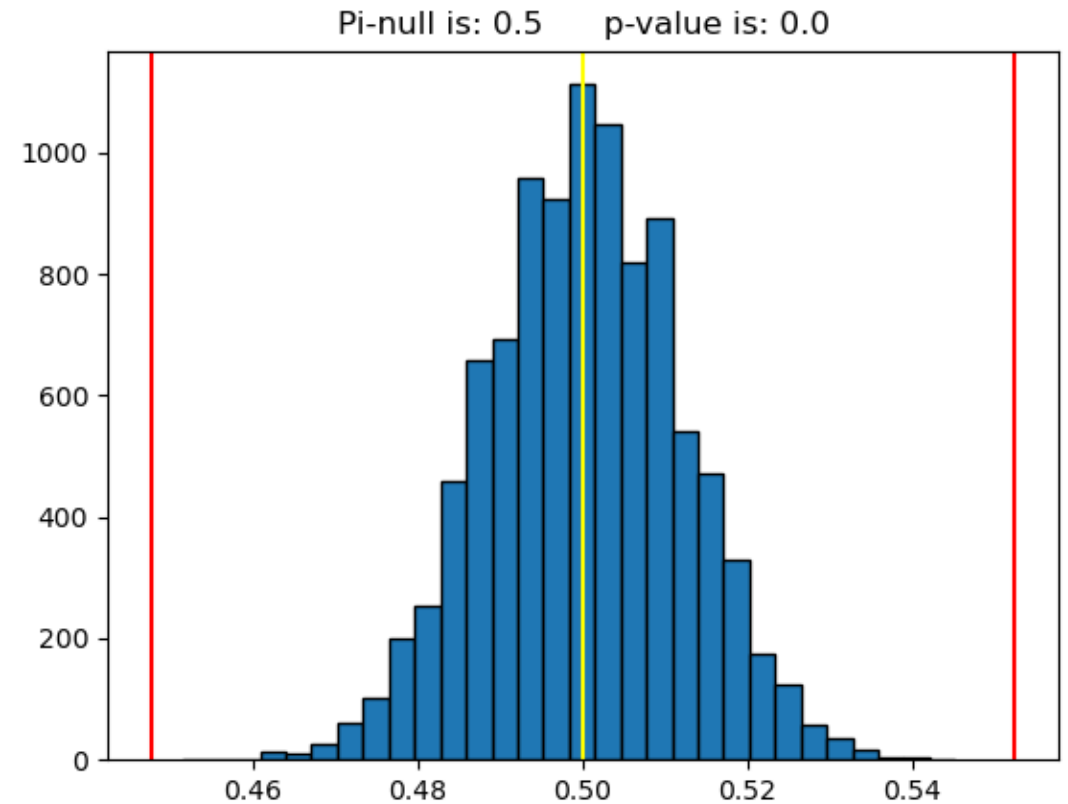
- $H_A: \pi_{\text{Bechdel}} \neq .5$
- $H_A: \mu_{\text{non-smoke}} \neq \mu_{\text{smoke}}$

Brief mention: two-sided hypothesis tests

When we have a “two-sided” alternative hypothesis, we need to calculate the the statistics that are “more extreme” than the observed statistic to get the p-value

- i.e., we need to look at both tails of our null distribution to get the p-value

Let's explore this in Jupyter!



Confidence intervals

Interval estimate based on a margin of error

Null hypothesis tests tell us if a particular parameter value is **implausible**

- E.g., in the Bechdel data we rejected $\pi = .5$

An **interval estimate** give a range of **plausible** values for a population parameter

Example: 42% of American approve of Biden's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter π lies somewhere between 39% to 45%

- i.e., if they sampled all voters the true population proportion would be likely be in this range

Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter

Think ring toss...

Parameter exists in the ideal world

We toss intervals at it

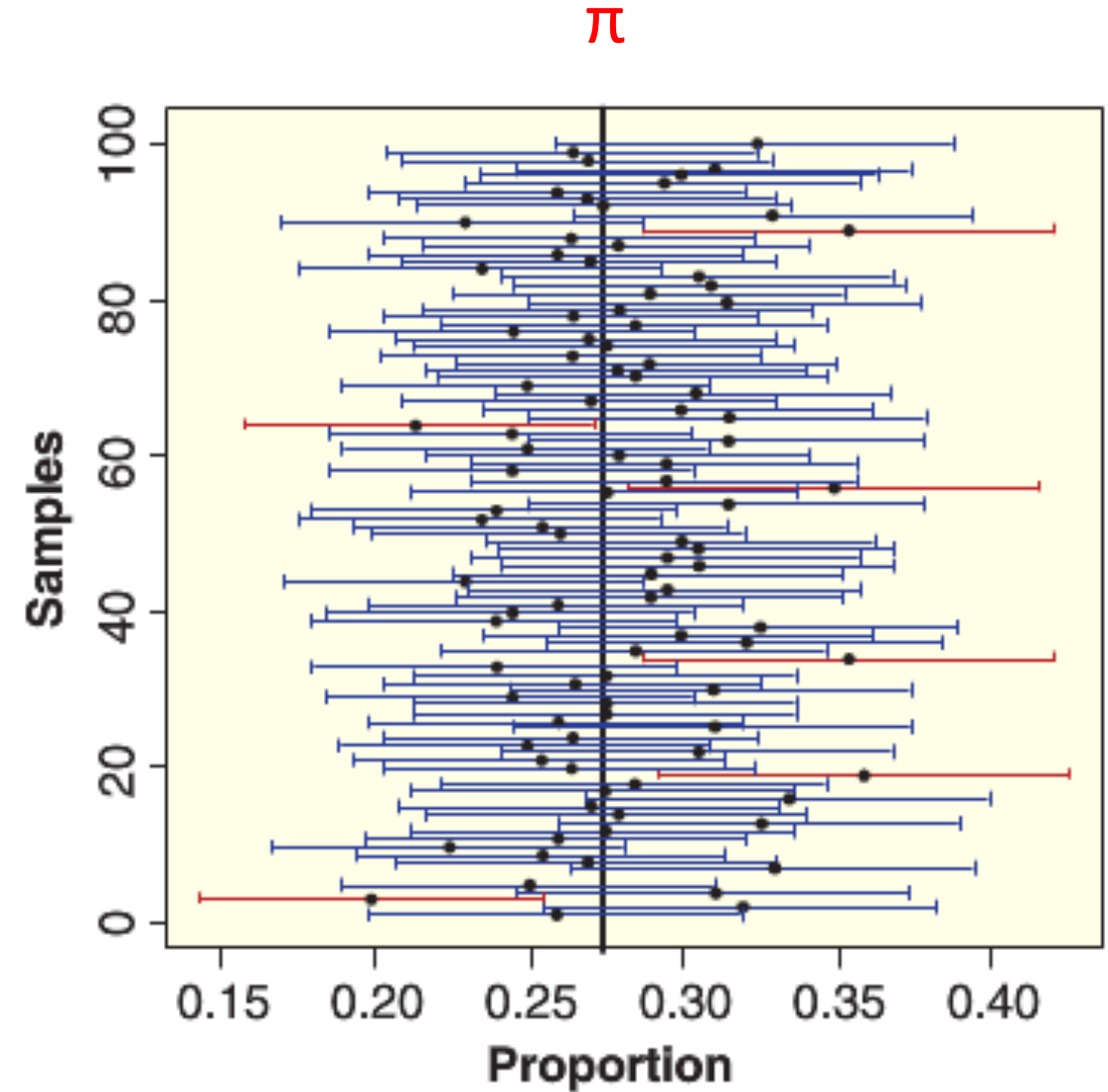
95% of those intervals capture the parameter



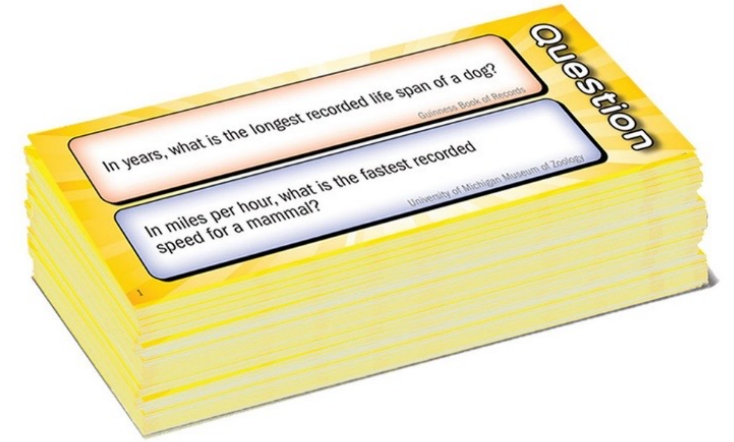
Confidence Intervals

For a **confidence level** of 95%...

95% of the **confidence intervals** will have the parameter in them



Wits and Wagers: 90% confidence interval estimator



I will ask 10 questions that have numeric answers

Please come up with a range of values that contains the true value in it for 9 out of the 10 questions

- i.e., be a 90% confidence interval estimator

Wits and Wagers...

Question 1: What is the diameter of the moon (in miles)?

Question 2: How many years passed between the first NBA game and the first WNBA game?

Question 3: What percent of U.S. land area does Alaska make up?

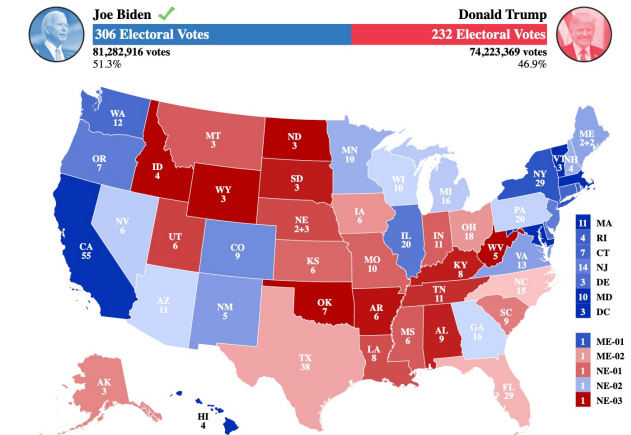
Wits and Wagers...

Question 4: On average, how many baseballs are used in a Major League Baseball season?

Question 5: How many rooms are there in the White House?

Question 6: How many votes were cast in the 2012 U.S. presidential election?

Question 7: Out of the 538 electoral votes, how many did Ronald Reagan receive in the 1984 presidential election ?



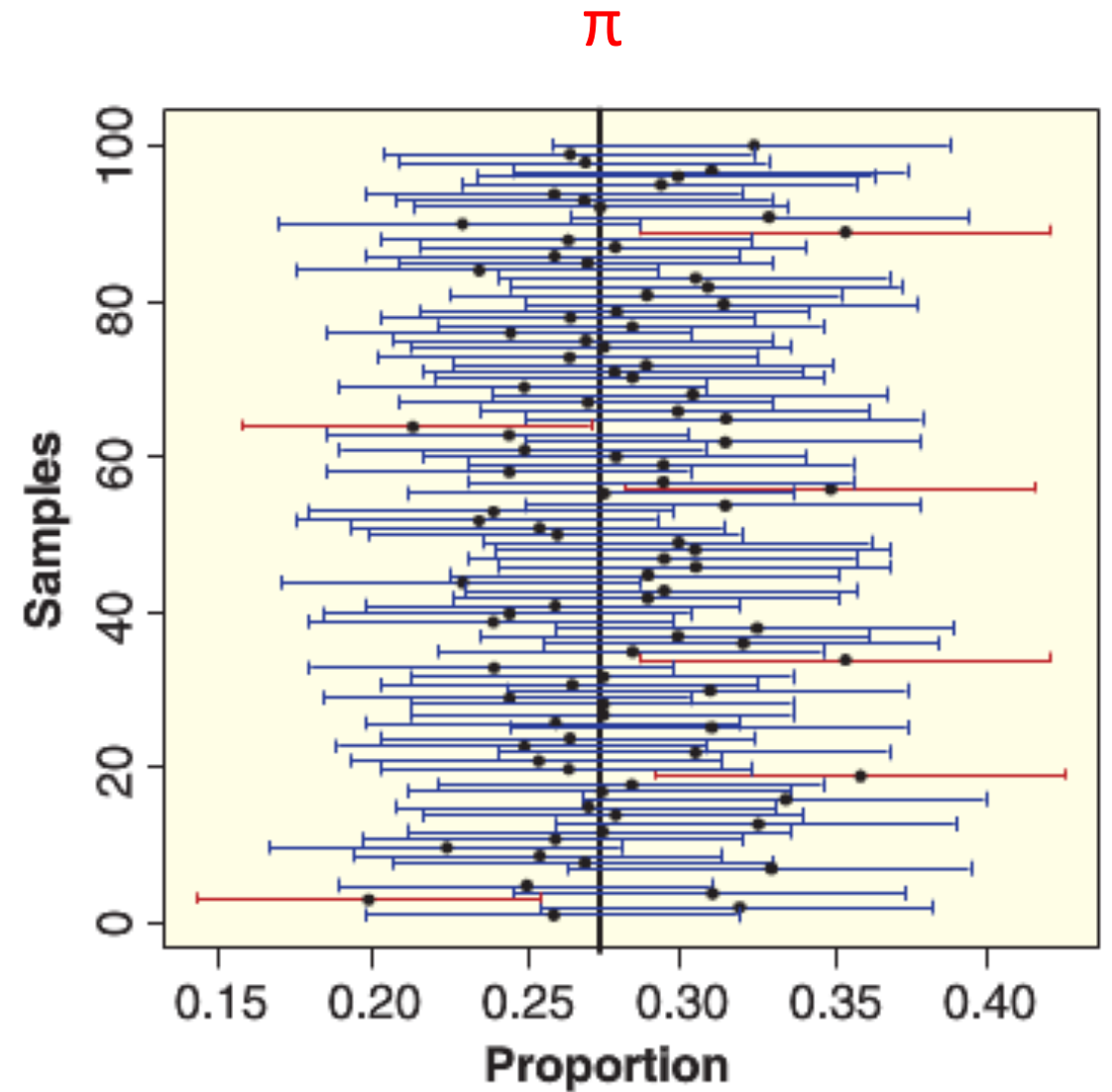
Wits and Wagers...

Question 8: How many cases of human spontaneous combustion appeared in medical journals between the years of 1600 and 1900?

Question 9: How many Academy Award nominations did *The Lord of the Rings* movie trilogy receive?

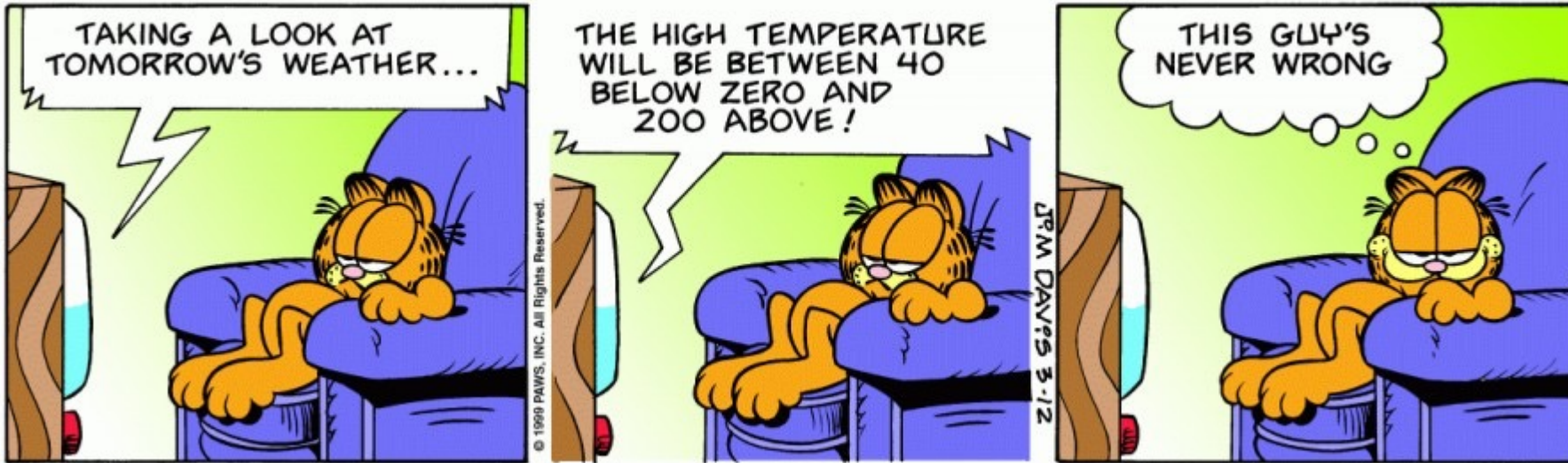
Question 10: In feet, how long was the largest whale ever recorded?

HOW DID WE DO?



We all have 9 out of 10 correct?!

Tradeoff between interval size and confidence level



There is a tradeoff between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

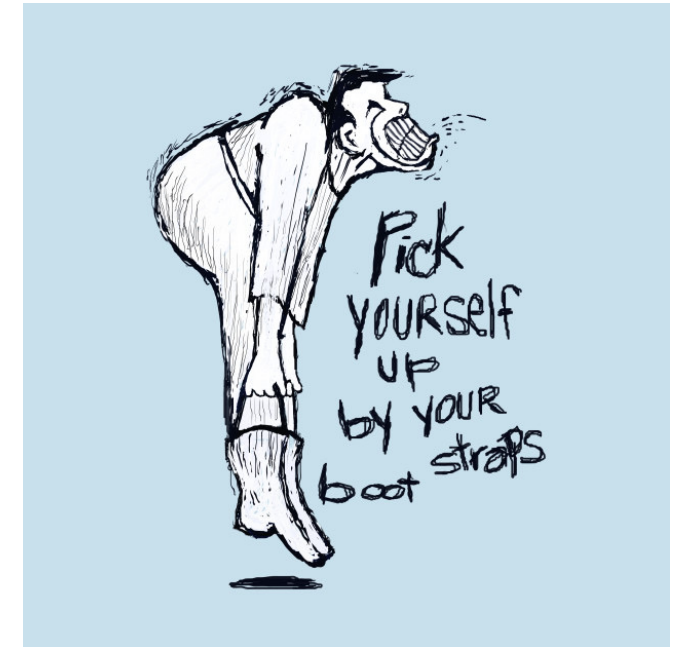
Using hypothesis tests to
construct confidence intervals

Constructing confidence intervals

There are several methods that can be used to construct confidence intervals including

- “Parametric methods” that use probability functions
 - E.g., confidence intervals based on the normal distribution
- A “bootstrap method” where data is resampled from our original sample to approximate a sampling distribution

To learn more about these methods, take Introductory Statistics!



Constructing confidence intervals

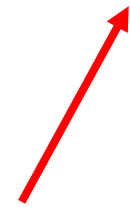
We are going to use a less conventional method to get confidence intervals based on the relationship between confidence intervals and hypothesis tests

- The method we will discuss is valid, but can be more computationally expensive than other methods

What we will do is to run a series of hypothesis test with different null hypothesis parameter values

Our confidence interval will be all parameters values where we **fail to reject** the null hypothesis

$$H_0: \pi = \pi_0$$



Failure to reject $\pi = \pi_0$
means π_0 is plausible

Motivation: Bechdel Confidence Interval

From running a hypothesis test on the Bechdel data, we saw that $H_0: \pi = .5$ is unlikely

- i.e., it was not plausible that 50% of movies pass the Bechdel test

But what is a reasonable range of values for the population proportion of movies that pass the Bechdel test?

Let's create a confidence interval for $H_0: \pi_{\text{Bechdel}}$ to find out!

Let's explore this in Jupyter!

