YData: Introduction to Data Science

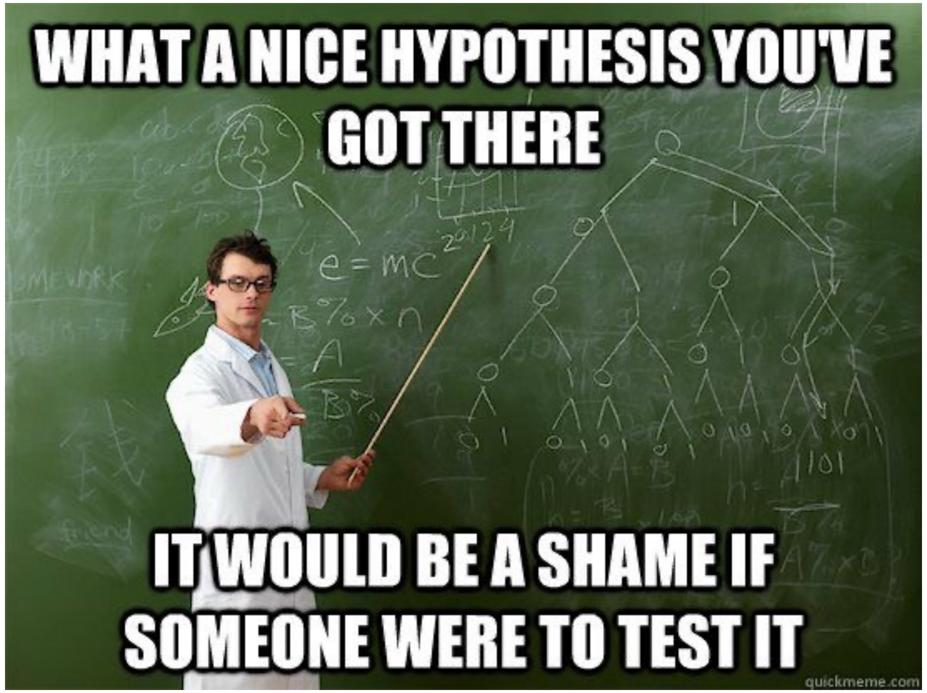
Class 21: Hypothesis tests for two means

Overview

Hypothesis test for two proportions and causality continued

Hypothesis tests for two means

If there is time: confidence intervals


WHAT A NICE HYPOTHESIS YOU'VE GOT THERE
IT WOULD BE A SHAME IF SOMEONE WERE TO TEST IT
quickmeme.com

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 manage 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

Review of Statistical Inference

Review: Statistical Inference

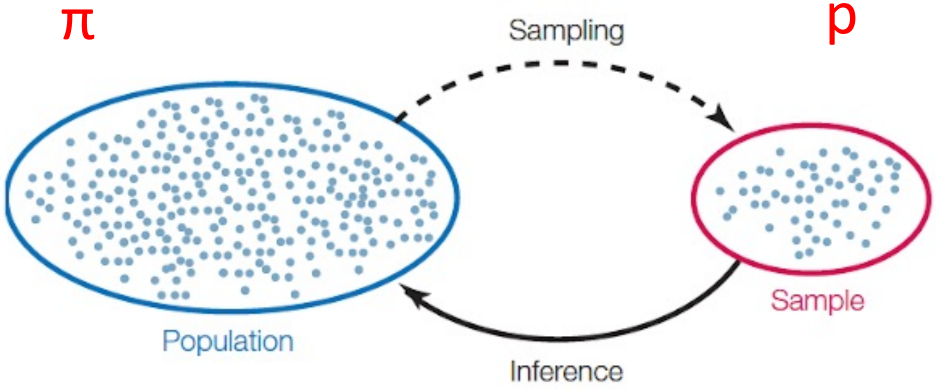
Statistical Inference: Making conclusions about a population based on data in a random sample

parameter

π

statistic

$\hat{p}$



A parameter is number associated with the population

- e.g., population proportion π
- e.g., the proportion of all voters who voted for Biden

A **statistic** is number calculated from the sample

- e.g., sample proportion $\hat{p}$
- e.g., the proportion of Biden's vote out of 1,000 people in a sample

A statistic can be used as an estimate of a parameter

	Sample Statistic	**Population Parameter**
Mean	$\bar{x}$	μ
Proportion	$\hat{p}$	π

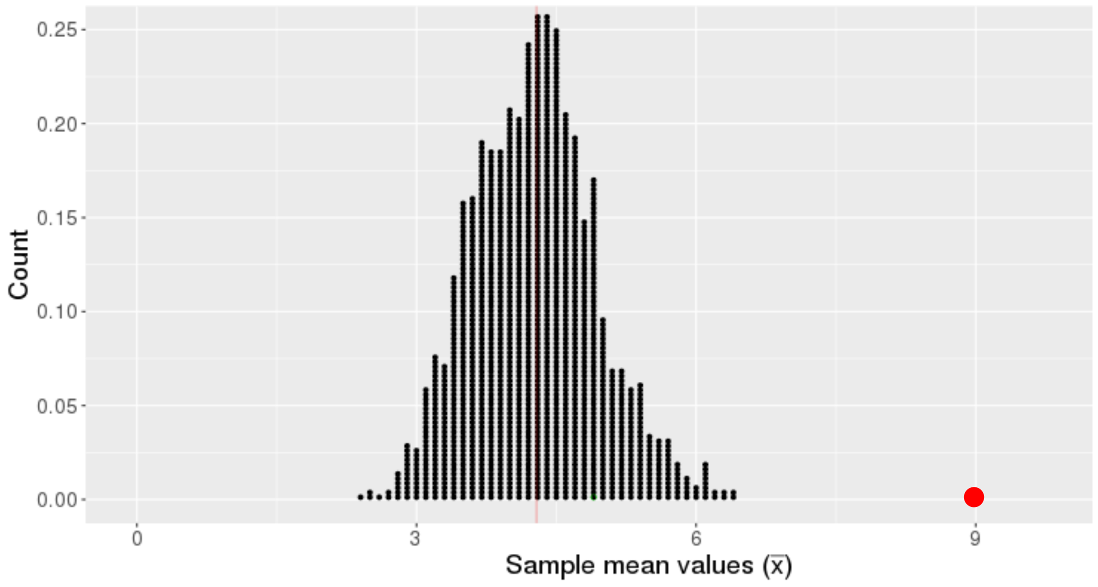
Hypothesis tests

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

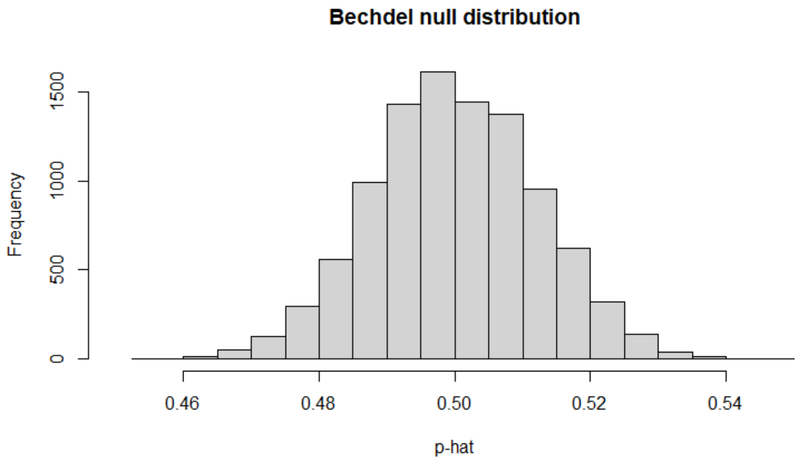


If our observed statistic is highly unlikely, we reject the claim

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

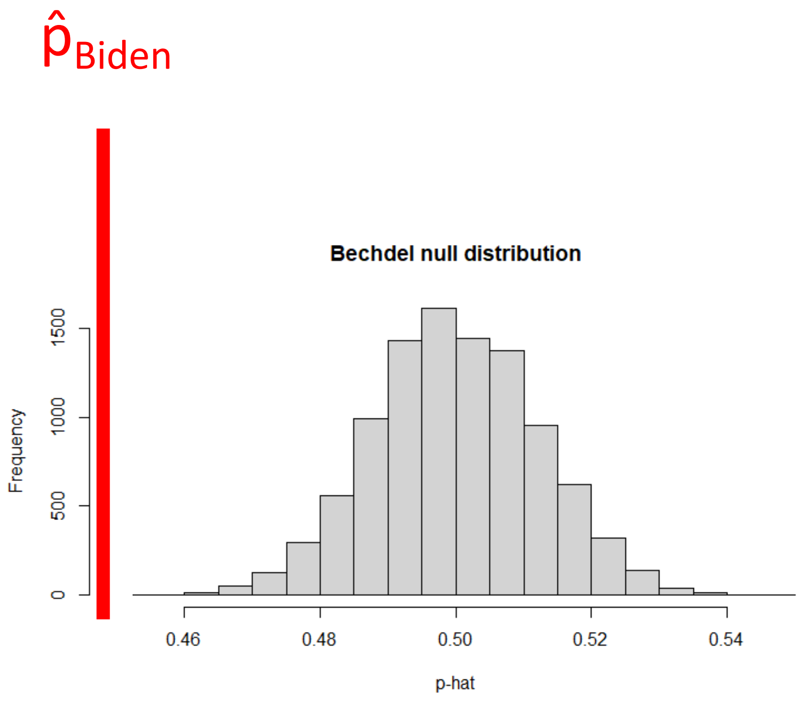
- The hypothesis we believe in (would like to see true)
- E.g., H_A: $\pi < 0.5$

A test statistic is the statistic we choose to simulate in order 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: = 0.5
- And we accept the alternative hypothesis
  - E.g., H_A: $\pi < 0.5$

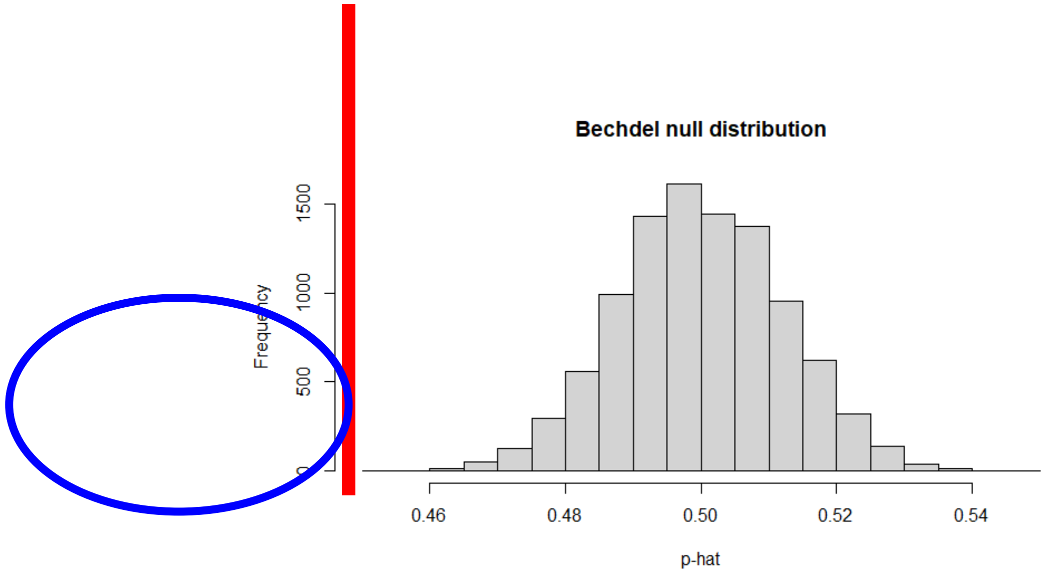


The p-value

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

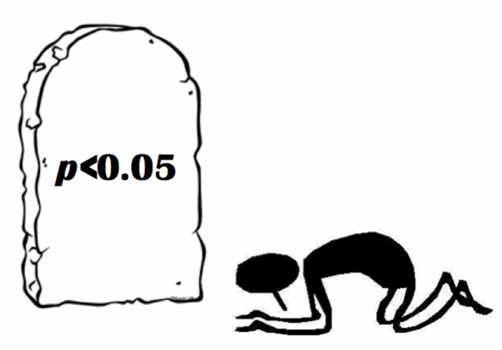
- P(Null_Stat ≤ obs_stat | H_0)

$\hat{p}_{\text{Biden}}$



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

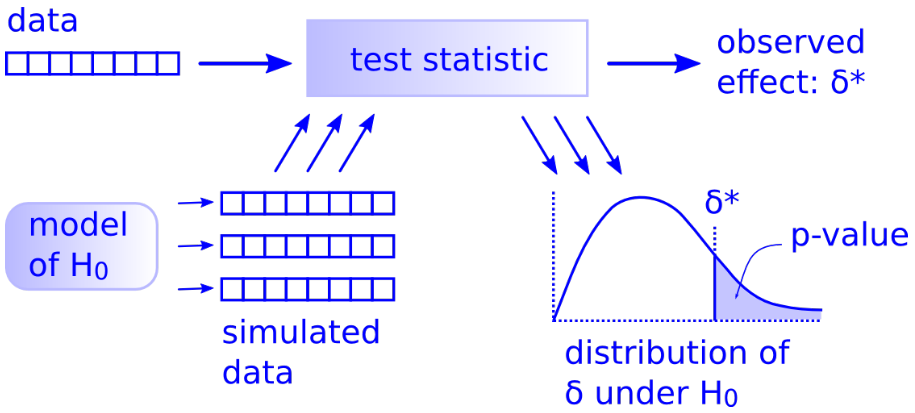
- Convention, p-value < 0.05

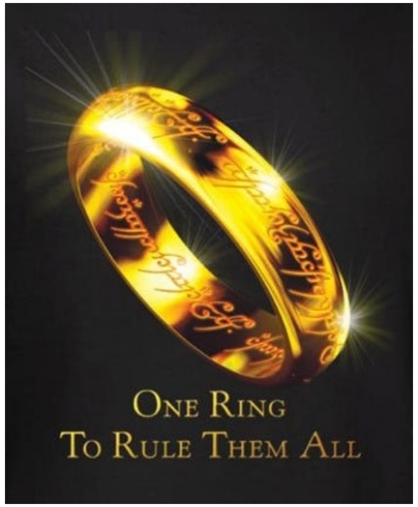


Steps needed to run a hypothesis test

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



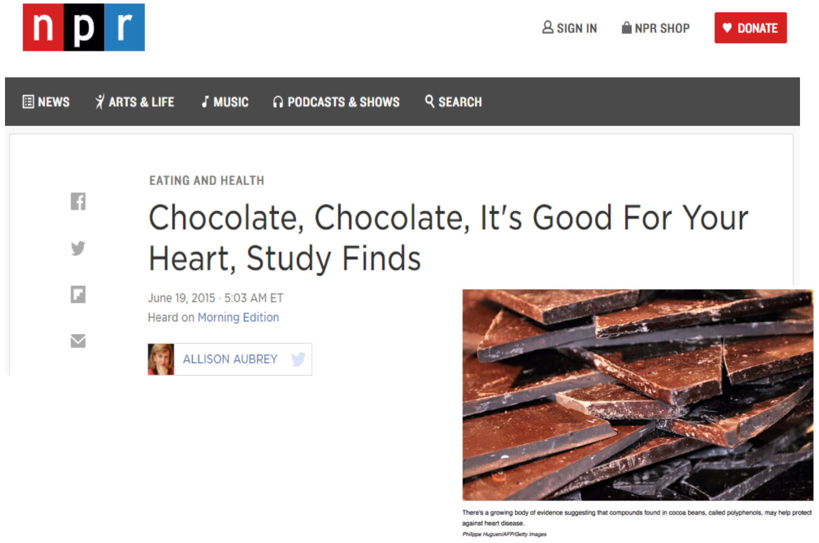


Assessing causal relationships

Review: Causality

Recall from class 2:

- **An association** is the presence of a reliable relationship between the treatments an outcome
- **A causal relationship** is when changing the value of a treatment variable influences the value outcome variable
- A **confounding variable** (also known as a **lurking variable**) is a third variable that is associated with both the treatment (explanatory) variable and the outcome (response) variable
  - A confounding variable can offer a plausible explanation for an association between the other two variables of interest



EATING AND HEALTH

Chocolate, Chocolate, It's Good For Your Heart, Study Finds

June 19, 2015 · 5:03 AM ET

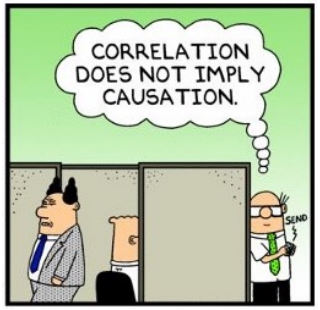
Heard on Morning Edition

ALLISON AUBREY

There's a growing body of evidence suggesting that compounds found in cocoa beans, called polyphenols, may help protect against heart disease.

Philippe Huguen/AFP/Getty Images

SOMEONE SENT ME ANOTHER ANONYMOUS EMAIL WITH A LINK TO AN ARTICLE ABOUT THE WORLD'S WORST BOSSES.
I GET ONE OF THOSE EMAILS EVERY TIME I LEAVE YOUR CUBICLE. DID YOU THINK I WOULDN'T NOTICE THE CORRELATION?



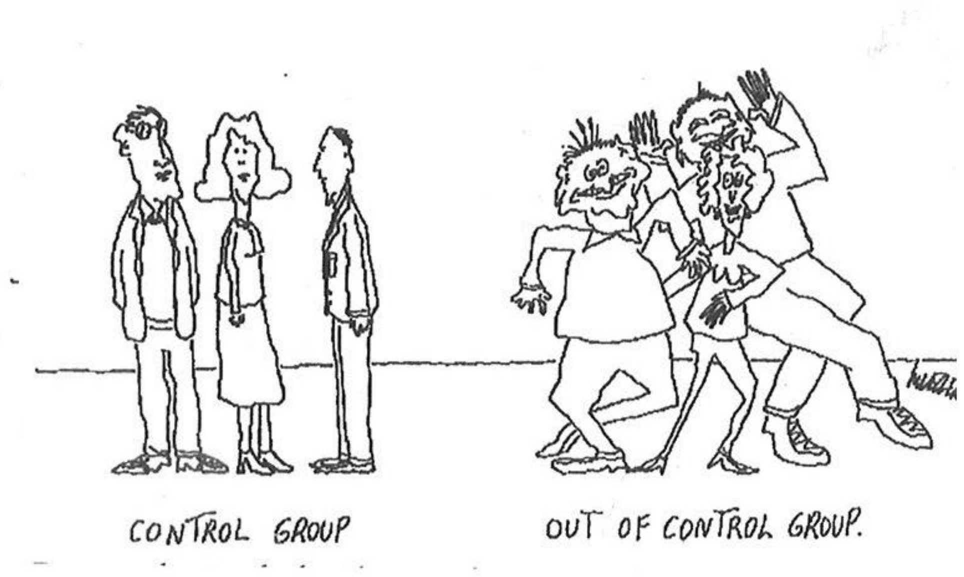
Lurking variable

Randomized Controlled Experiment

Sample A: control group

Sample B: treatment group

If members of the treatment and control groups are selected at random; this allows causal conclusions!



In particular, any difference in outcomes between the two groups could be due to:

- Chance
- The treatment

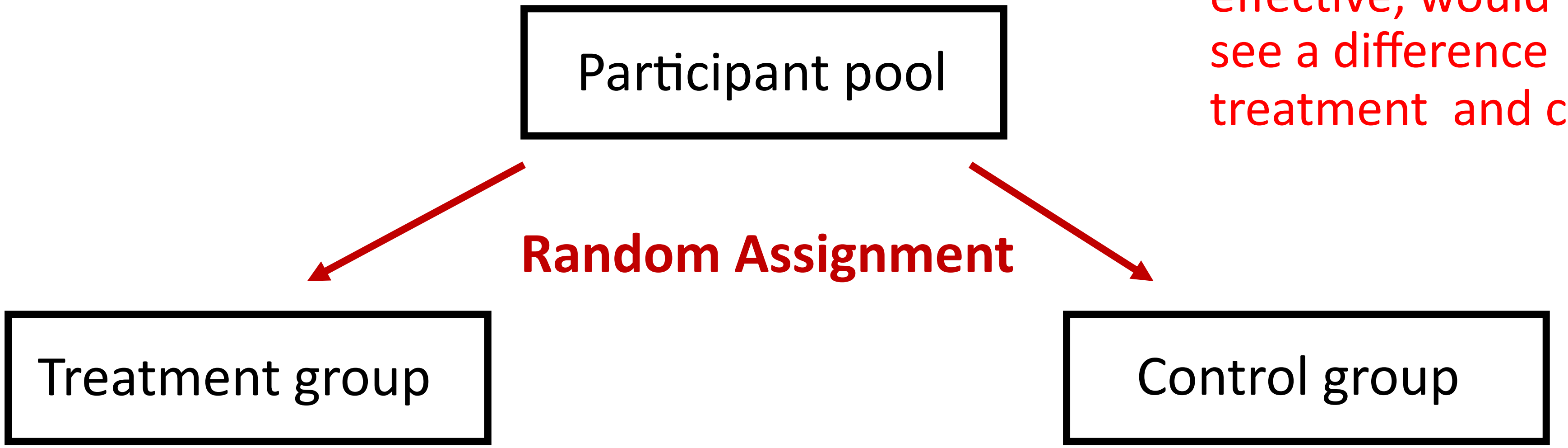
Randomly assigning participants to treatment and control groups allows us to separate what expected by chance and consequently what is due to the treatment

Randomized Controlled Experiment

Take a group of participant and ***randomly assign***:

- Half to a *treatment group* where they get chocolate
- Half in a *control group* where they get a fake chocolate (placebo)
- See if there is more improvement in the treatment group compared to the control group

Q: If the treatment was not effective, would we expect to see a difference between the treatment and control groups?



Participant pool

Random Assignment

Treatment group

Control group

Case study

RCT to study Botulinum Toxin A (BTA) as a treatment to relieve chronic back pain

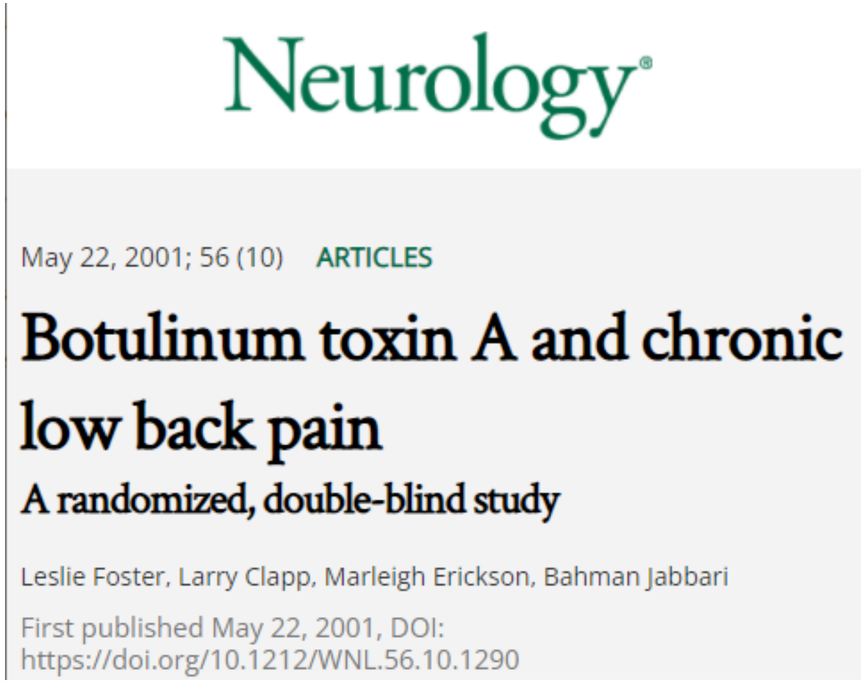
- 15 patients in the treatment group (received BTA)
- 16 in the control group (normal saline)

Trials were run double-blind: neither doctors nor patients knew which group they were in.

Results

- 2 patients in the control group had relief from pain (outcome=1)
- 9 patients in the treatment group had relief.

Can this difference be just due to chance?



Neurology®

May 22, 2001; 56 (10) ARTICLES

Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

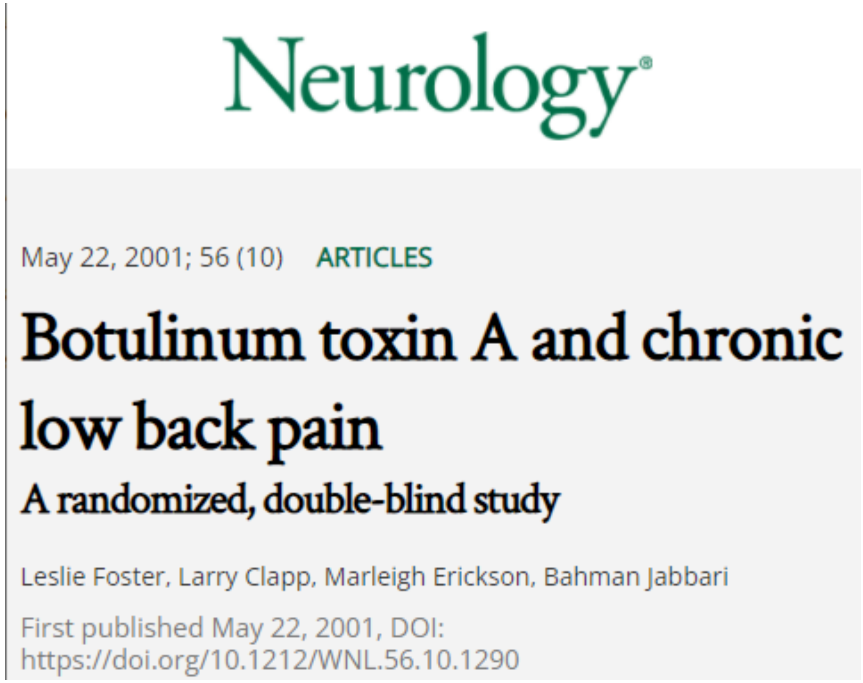
Step 1: The hypotheses

Null:

- BTA does not lead to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be the same in both groups.
  - H_0: $\pi_{treat} = \pi_{control}$

Alternative:

- BTA leads to an increase in pain relief
  - i.e., if many people were to get BTA and saline, the proportion of people who experienced pain relief would be higher for those who received BTA
  - H_A: $\pi_{treat} > \pi_{control}$



Neurology®

May 22, 2001; 56 (10) ARTICLES

Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

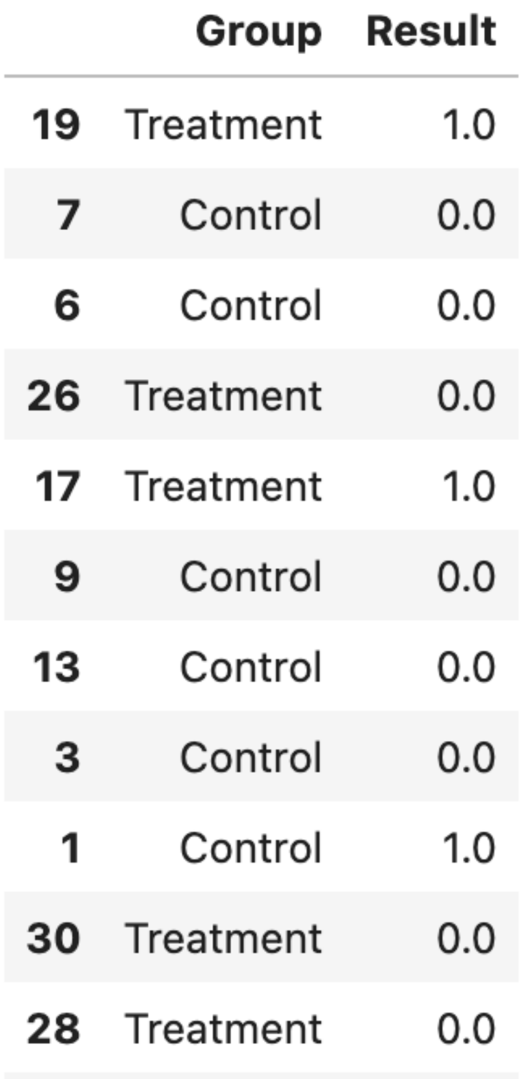
Step 2: The observed statistic

To calculate an observed statistic we need data:

Let’s have our observed statistic mirror our hypotheses

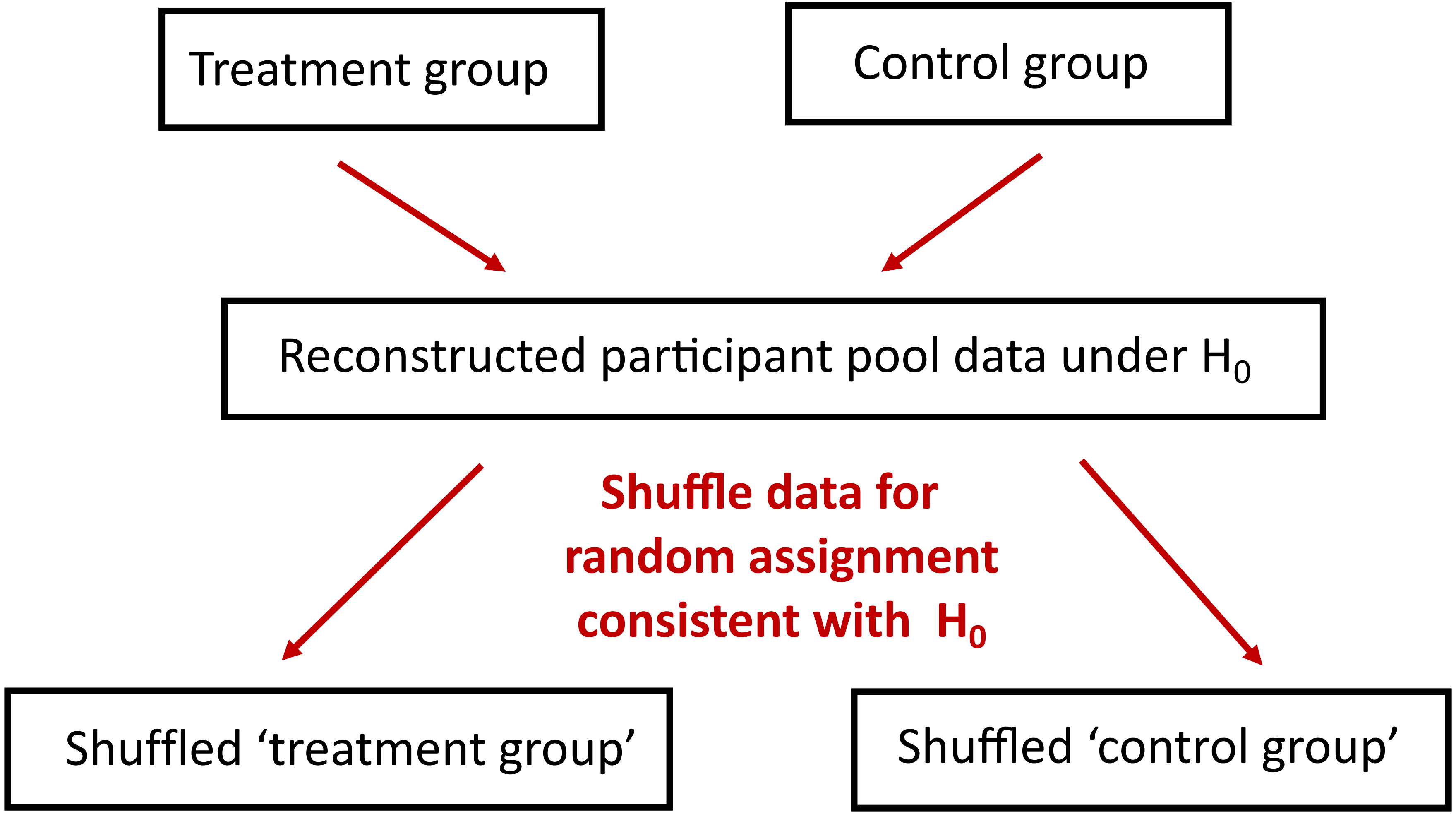
- H_0: $\pi_{treat} - \pi_{control} = 0$

Observed statistic is: $\hat{p}_{treat} - \hat{p}_{control}$



	Group	Result
19	Treatment	1.0
7	Control	0.0
6	Control	0.0
26	Treatment	0.0
17	Treatment	1.0
9	Control	0.0
13	Control	0.0
3	Control	0.0
1	Control	1.0
30	Treatment	0.0
28	Treatment	0.0

3. Create the null distribution!



Treatment group

Control group

Reconstructed participant pool data under H_0

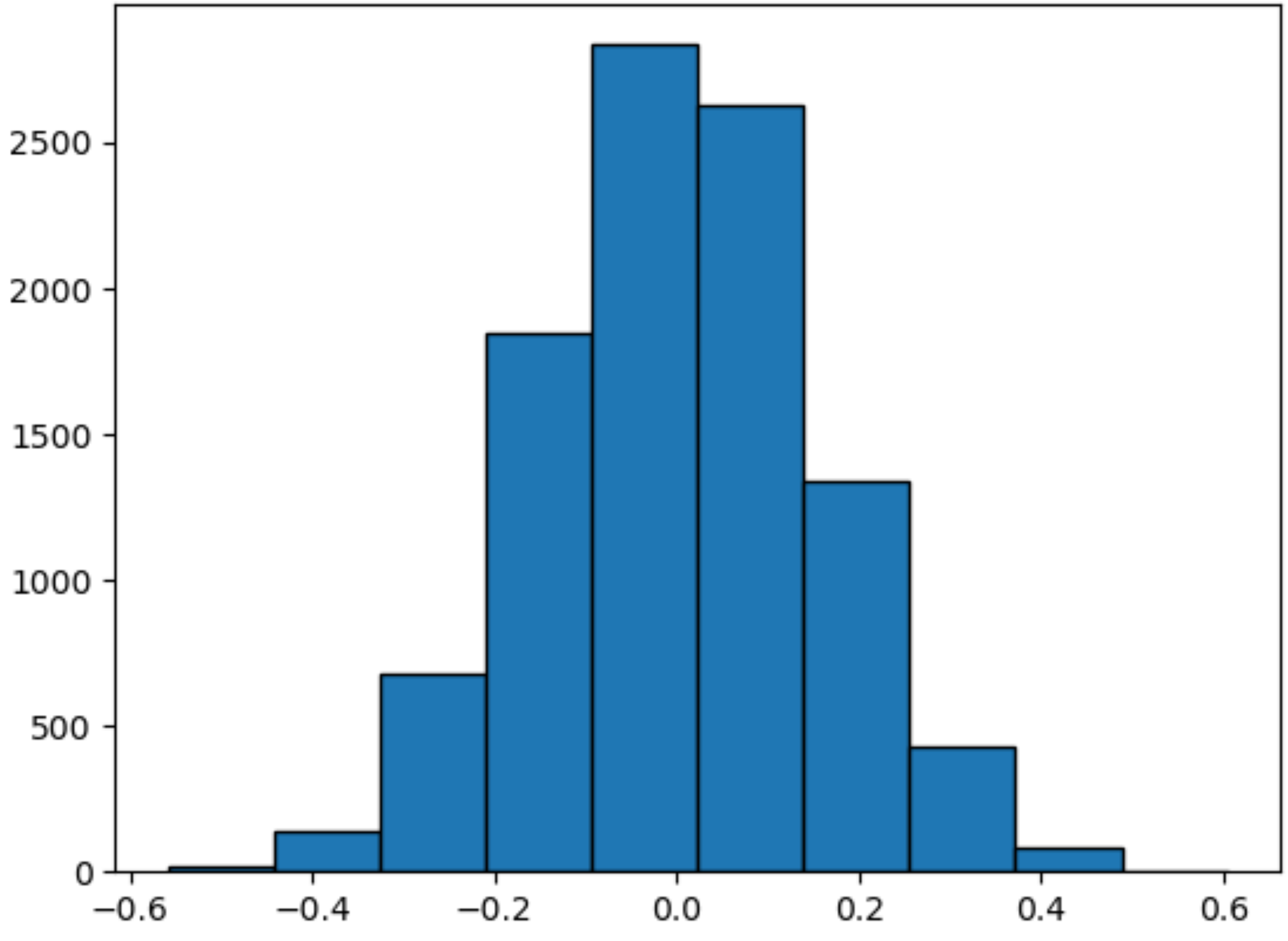
Shuffle data for random assignment consistent with H_0

Shuffled 'treatment group'

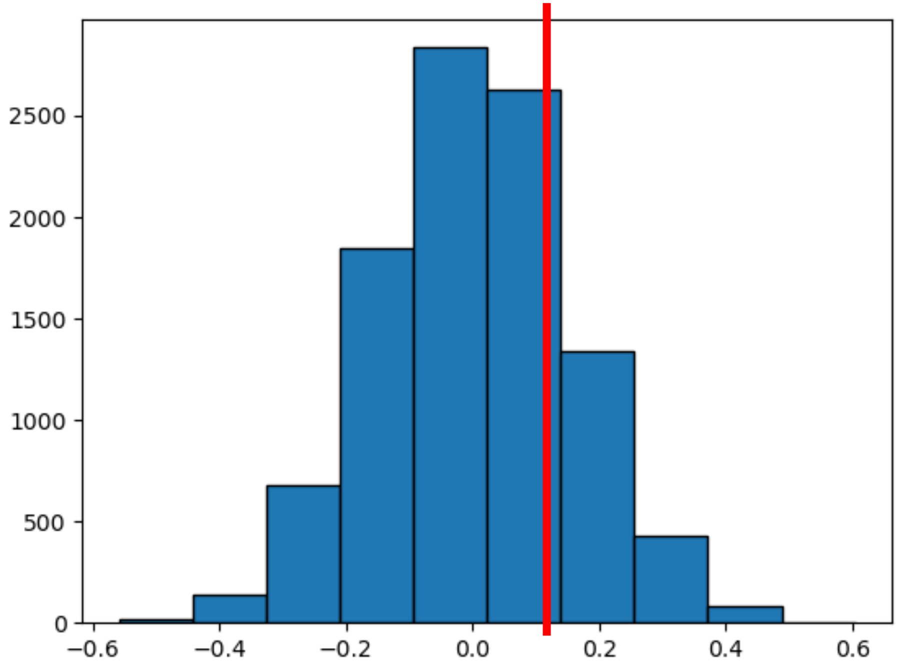
Shuffled 'control group'

One null distribution statistic: $\hat{p}_{Shuff\_Treatment} - \hat{p}_{Shuff\_control}$

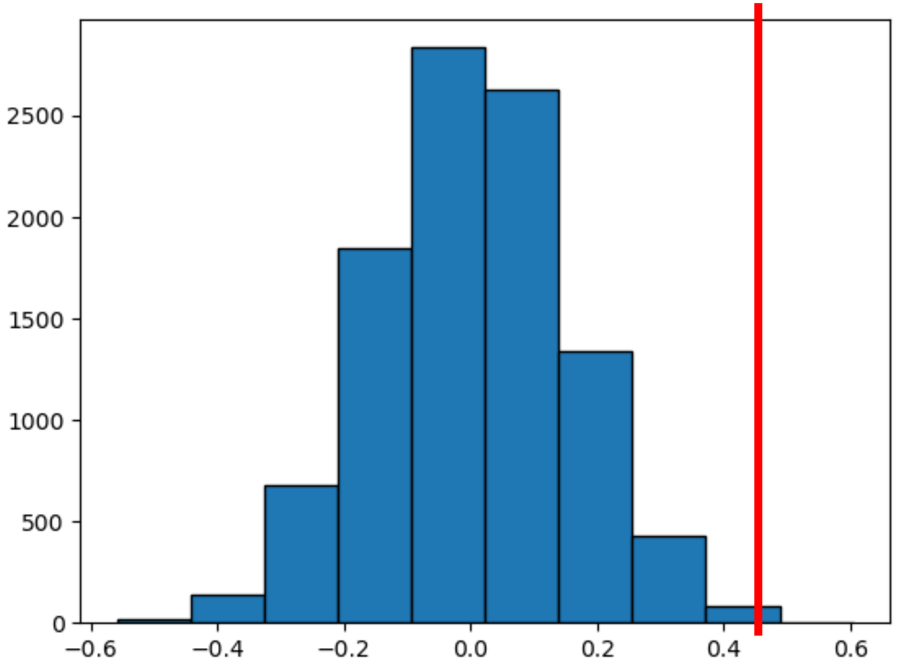
Step 3: Create a null distribution



Step 4: Calculate the p-value

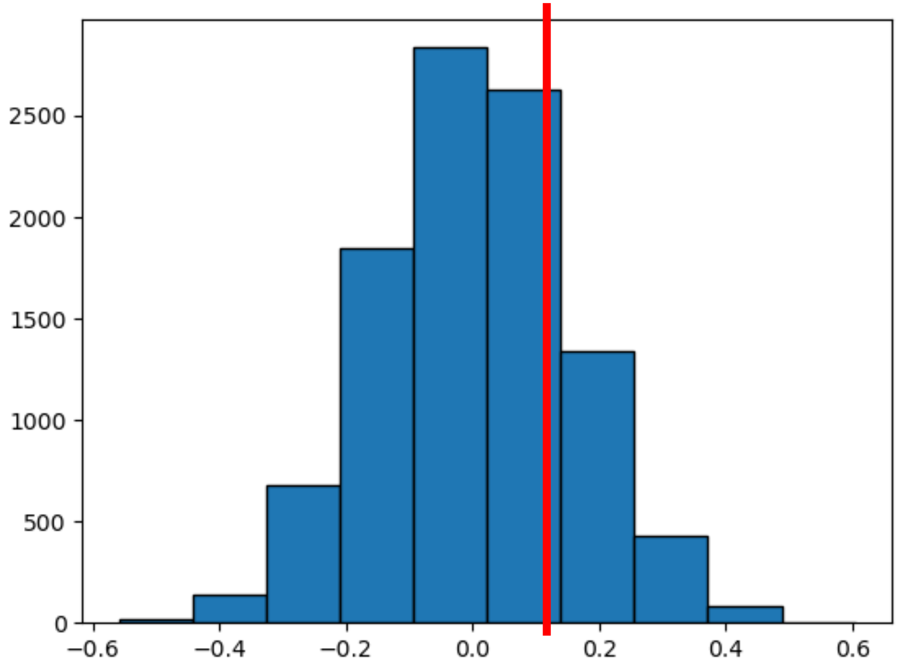


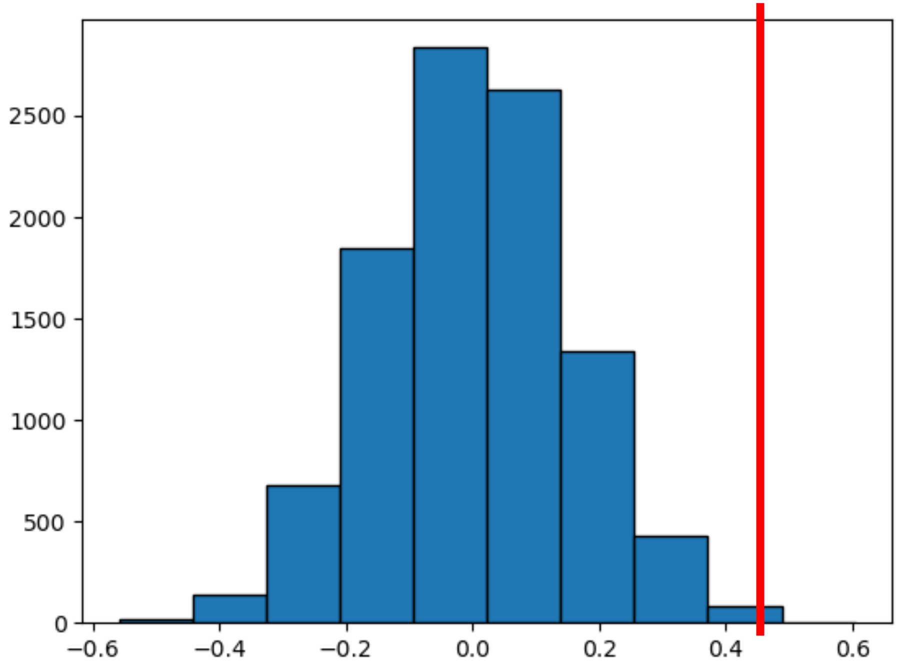
If $\hat{p}_{treat}$ - $\hat{p}_{control}$ = 0.1 what would the p-value be?



If $\hat{p}_{treat}$ - $\hat{p}_{control}$ = 0.5 what would the p-value be?

Step 5: Draw a conclusion





If the p-value was 0.19 what would we conclude?

If the p-value was 0.0007 what would we conclude?

JANUARY 16, 2017

TIME

Depression.
Heart trouble.
Migraines.
Erectile dysfunction.
Back pain.
Sweaty palms.
Drooling.
And 793 other problems.

How Botox Became the Drug That's Treating Everything.

By Alexandra Sifferlin

time.com

Let's explore this in Jupyter!

Summary: BTA for back pain relief

1. State the null hypothesis and the alternative hypothesis

- BTA does not lead to an increase in pain relief: H_0: $\pi_{treat} = \pi_{control}$
- BTA leads to an increase in pain relief: H_A: $\pi_{treat} > \pi_{control}$

2. Calculate the observed statistic: $\hat{p}_{treat}$ - $\hat{p}_{control}$

3. Create a null distribution that is consistent with the null hypothesis

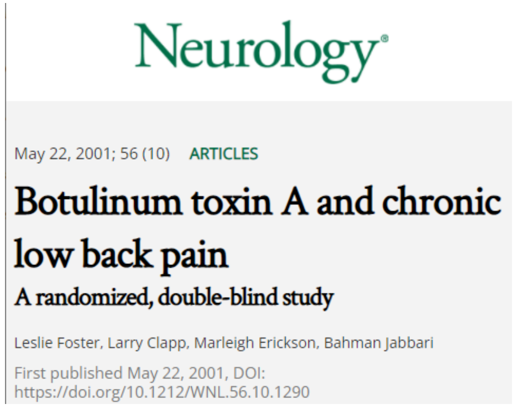
- The $\hat{p}_{treat}$ - $\hat{p}_{control}$ statistics we expect if the null hypothesis was true
- i.e., statistics we would expect if there was no difference in pain relief between the two groups

4. Examine how likely the observed statistic is to come from the null distribution

- What is the probability that we would get a $\hat{p}_{treat}$ - $\hat{p}_{control}$ statistic larger than 0.475 if the null hypothesis was true?
- i.e., what is the p-value?

5. Make a judgement

- A small p-value this means that at the proportion of pain relief differed between the two groups
  - i.e., we say our results are 'statistically significant'
- Because our analysis is based on a randomized controlled trial (using random assignment) we can say that BTA ***causes*** an increase in pain relief



Neurology®

May 22, 2001; 56 (10) ARTICLES

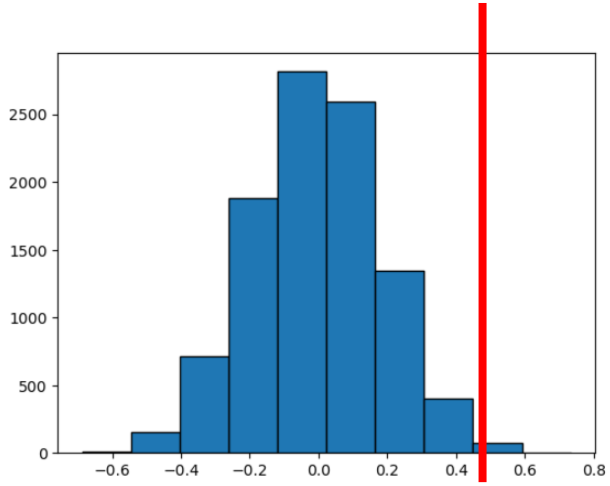
Botulinum toxin A and chronic low back pain

A randomized, double-blind study

Leslie Foster, Larry Clapp, Marleigh Erickson, Bahman Jabbari

First published May 22, 2001, DOI: https://doi.org/10.1212/WNL.56.10.1290

$\hat{p}_{treat}$ - $\hat{p}_{control}$ = .475



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

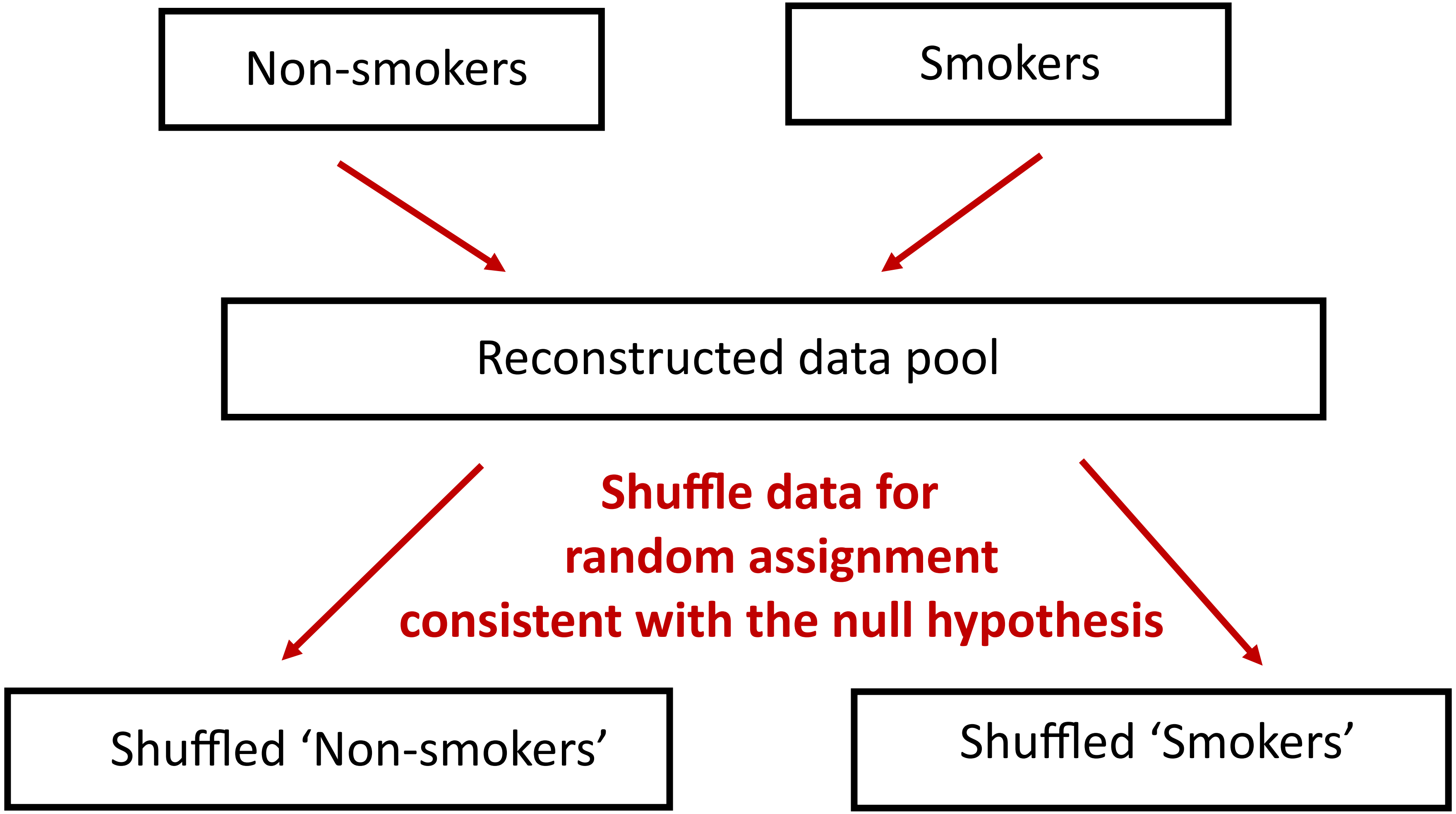
Step 3: Create the null distribution

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!



Non-smokers

Smokers

Reconstructed data pool

Shuffle data for random assignment consistent with the null hypothesis

Shuffled 'Non-smokers'

Shuffled 'Smokers'

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

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 π = .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

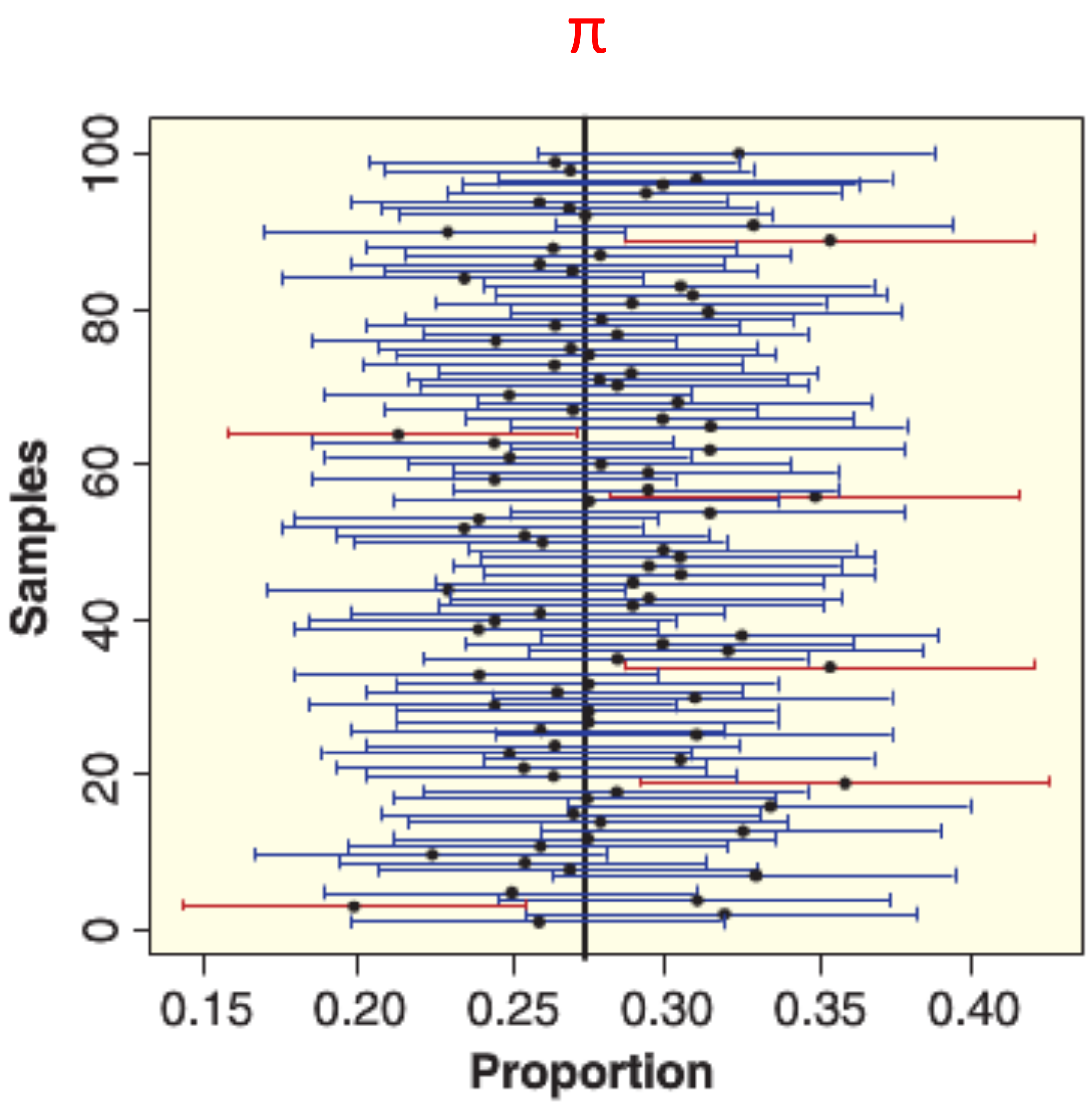
We toss intervals at it

95% of those intervals capture the parameter

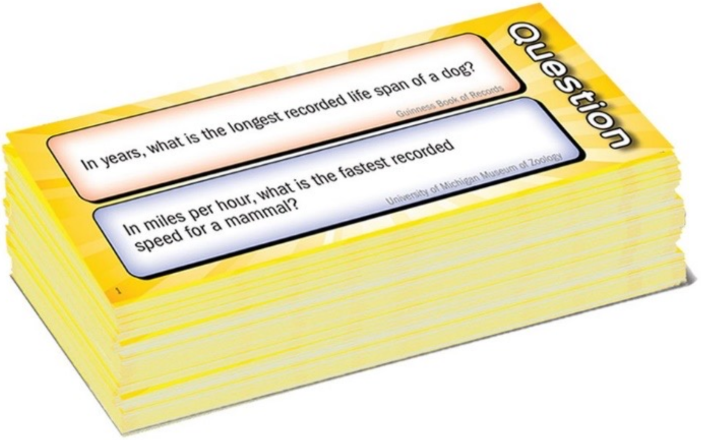
Confidence Intervals

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

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



Wits and Wagers:
90% confidence interval estimator


Question
In years, what is the longest recorded life span of a dog?
Guinness Book of Records
In miles per hour, what is the fastest recorded speed for a mammal?
University of Michigan Museum of Zoology
1

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?

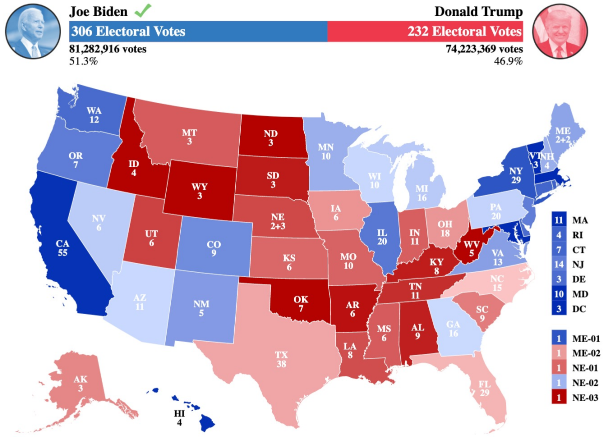
Wits and Wagers…

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…

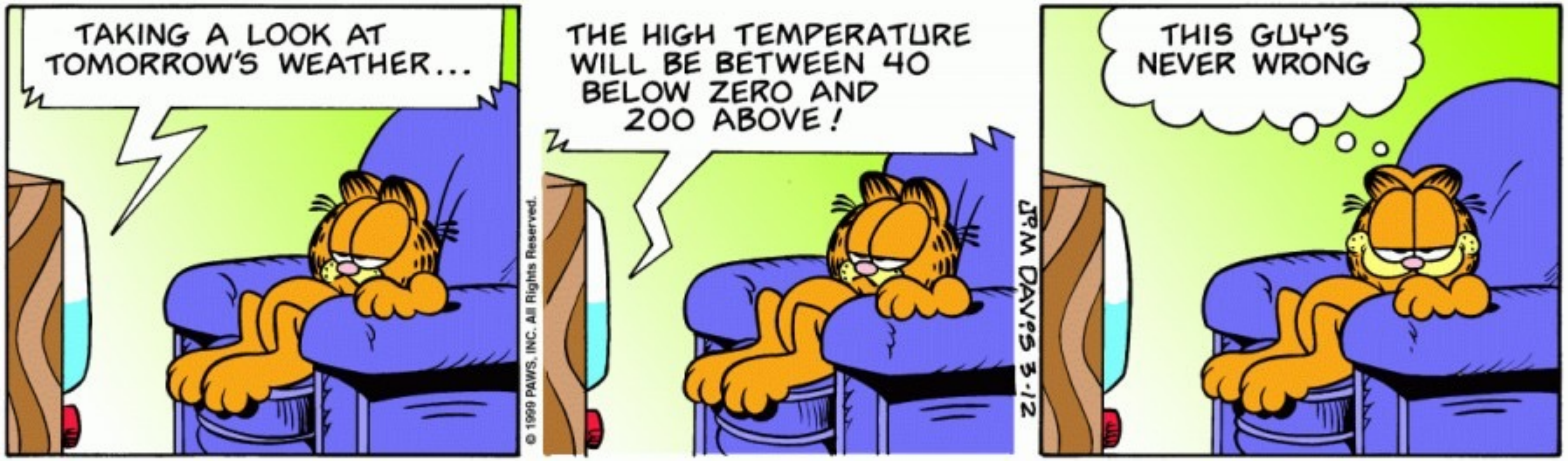
Question 8: How many cases of human spontaneous combustion appeared in medical journals between the years of 1600 and 1900?

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?

Tradeoff between interval size and confidence level


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JPM DAVIS 3-12

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

$H_0: \pi = \pi_0$

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

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

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

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

WELL,... IDUNNO. I HAVE THIS RULE, SEE...
I ONLY GO TO A MOVIE IF IT SATISFIES THREE BASIC REQUIREMENTS. ONE, IT HAS TO HAVE AT LEAST TWO WOMEN IN IT..
THE BARBARIAN
..WHO, TWO, TALK TO EACH OTHER ABOUT, THREE, SOMETHING BESIDES A MAN.
THE VIGILANTE
PRETTY STRICT, BUT A GOOD IDEA.
NO KIDDING. LAST MOVIE I WAS ABLE TO SEE WAS ALIEN...

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: π_{Bechdel} to find out!

Let's explore this in Jupyter!