YData: Introduction to Data Science
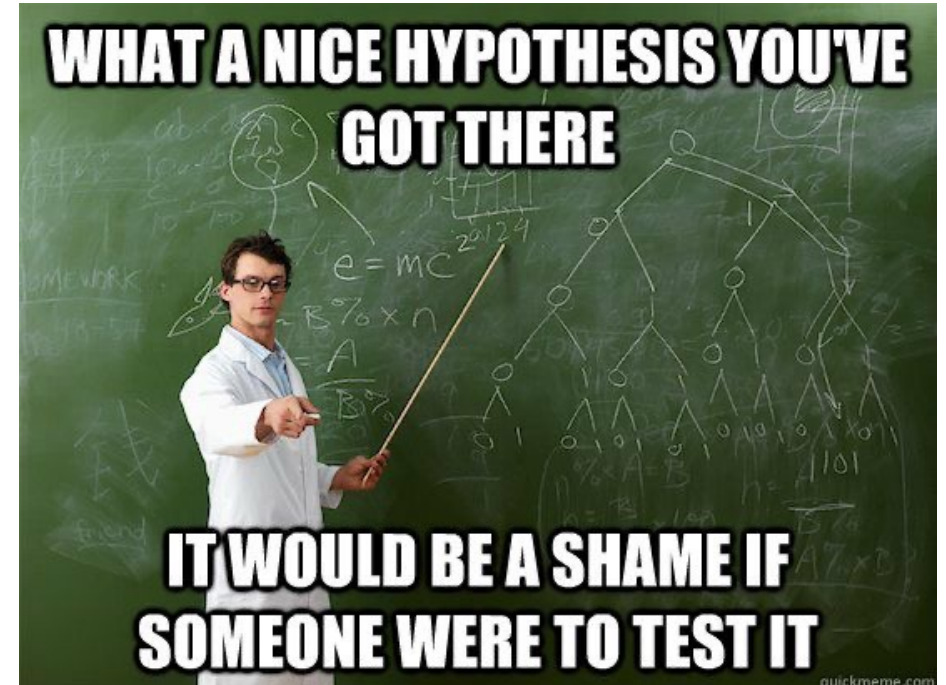

Class 21: Hypothesis tests continued

Overview

Quick review of parameters, statistics, sampling, and hypothesis tests for a single proportion

Hypothesis tests for multiple proportions

Hypothesis tests assessing causality

Project timeline

That went as planned, said no project ever.

Tuesday, April 11th
- Projects are due on Gradescope at 11pm on
- Also, email a pdf of your project to your peer reviewers
  - A list of whose paper you will review has been posted to Canvas

Wednesday, April 19th
- Jupyter notebook files with your reviews need to be sent to the authors and a pdf need to be submitted to Gradesscope
- A template for doing your review is available on Canvas

Sunday, April 30th
- Project is due on Gradescope
  - Add peer reviews to an 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 19th 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 Zihe (zihe.zheng@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 16th
- Thanks to Rose, it is not too long

Review of Statistical Inference

Review: Statistical Inference

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

A parameter is number associated with the population
- e.g., population proportion π
- e.g., the proportion of 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 our sample

A statistic can be used as an estimate of a parameter

parameter   statistic

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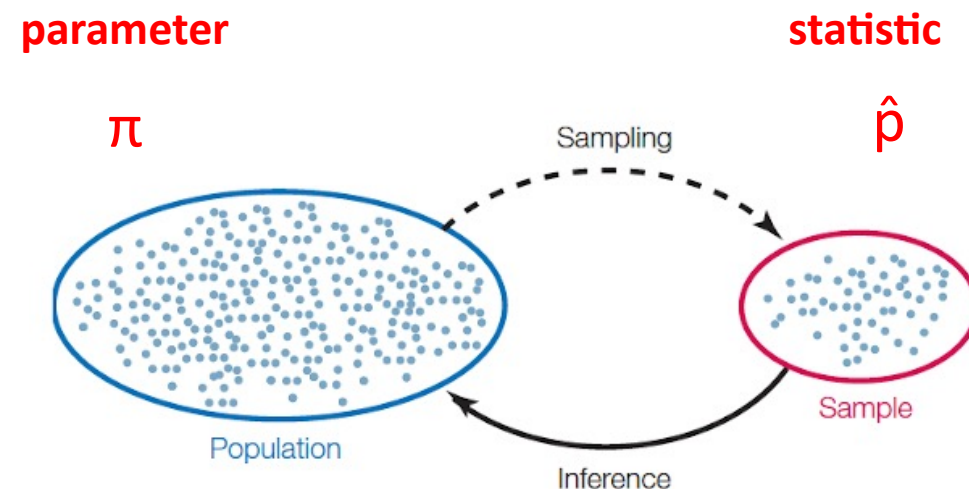
	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Proportion	$\hat{p}$	π

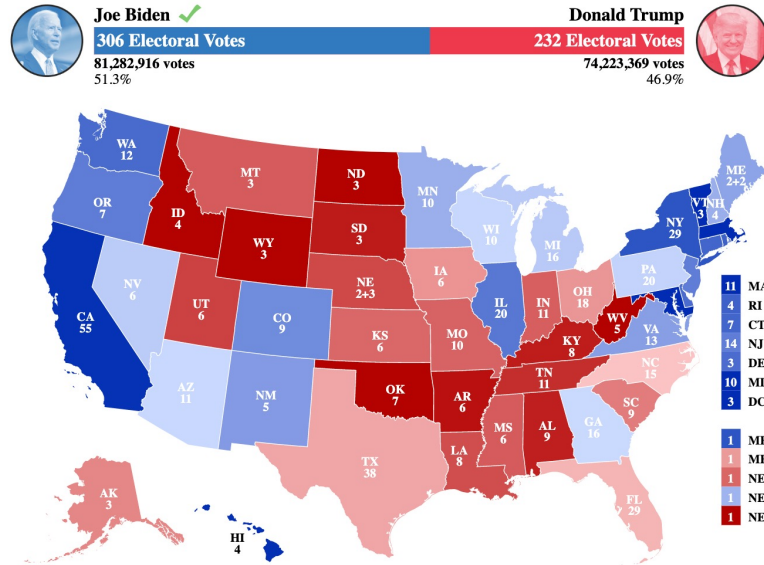# Probability distribution of a statistic

Values of a statistic vary because random samples vary

A **sampling distribution** is a probability distribution of *statistics*
- All possible values of the statistic and all the corresponding probabilities
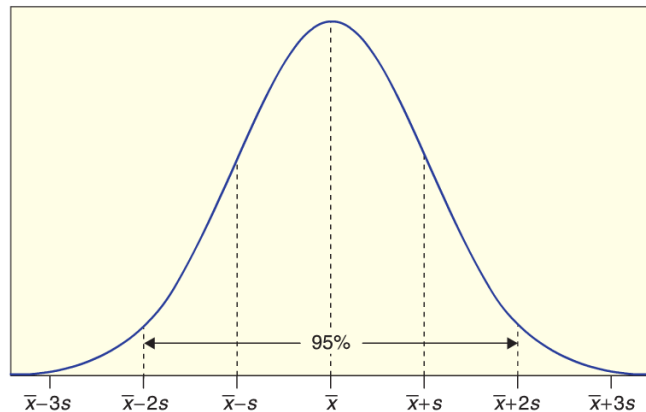- We can approximate a sampling distribution by simulating statistics

π_{Biden}

n = 1,000

$\hat{p}_{Biden}$

$\hat{p}_{Biden}$

$\hat{p}_{Biden}$

95%

$\overline{x}-3s$ $\overline{x}-2s$ $\overline{x}-s$ $\overline{x}$ $\overline{x}+s$ $\overline{x}+2s$ $\overline{x}+3s$

Sampling distribution!

Simulating random proportions (p̂'s)

We can simulate random proportions p̂ consistent with a population proportion π by:

1. Generated n random numbers uniformly distributed between 0 and 1
   - rand_nums = np.random.rand(1000)

2. Marking points less than π as being True, and greater π than as being False
   - rand_binary = rand_nums <= pi_value

3. Calculating the proportion of points to get a p̂
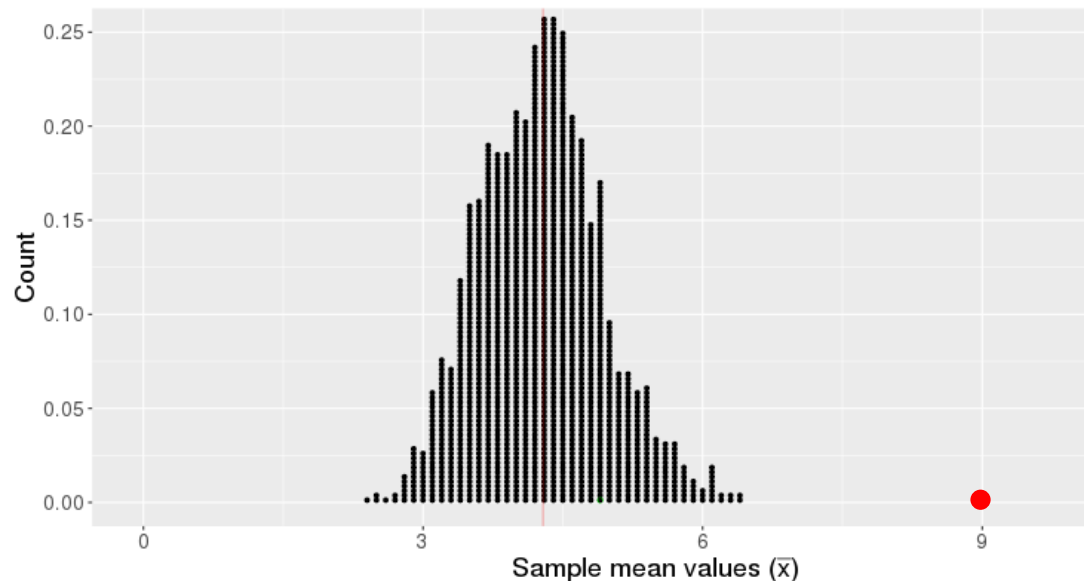   - rand_phat = np.mean(rand_binary)

π = .7

True   False

p̂

Hypothesis tests

Basic hypothesis test logic

We start with a claim about a population parameter
- E.g., μ = 4

This claim implies we should get a certain distribution of statistics

If our observed statistic is highly unlikely, we reject the claim
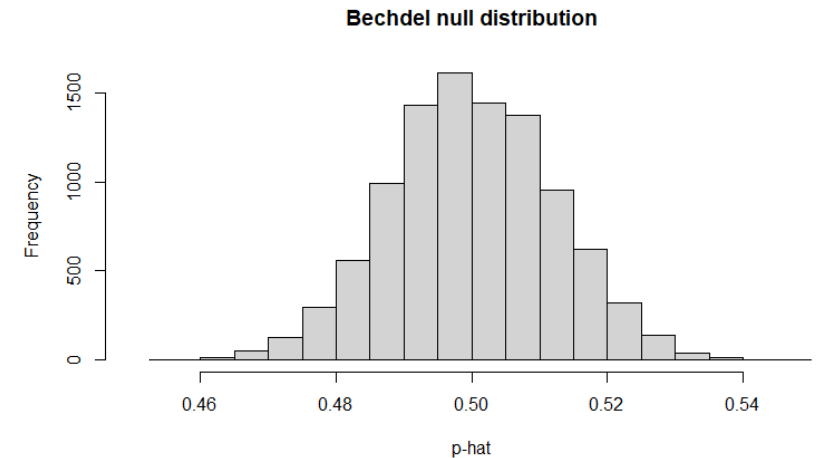
Null and Alternative hypotheses

Null hypothesis

- A hypothesis where "nothing interesting" happened
  - E.g., our experiment failed
- We can simulate data under the assumptions of this model to get a "null distribution" of statistics

Bechdel null distribution

Alternative hypothesis

- The hypothesis we believe in (would like to see true)

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

 - And we accept the alternative hypothesis

Bechdel null distribution
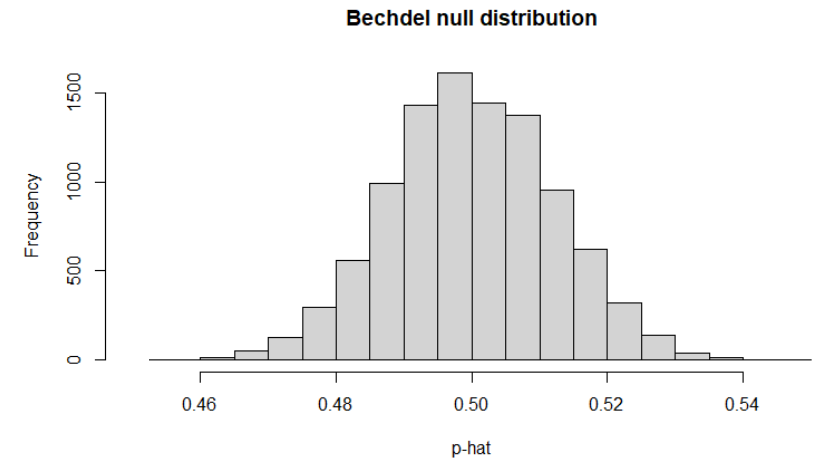
The p-value

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

- $P(\text{Null\_Stat} \geq \text{obs\_stat} \mid H_0)$

If the P-value is small, this is evidence against the null hypothesis and the results are often called "statistically significant"
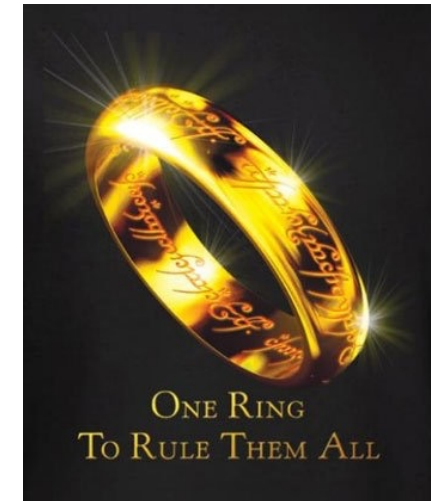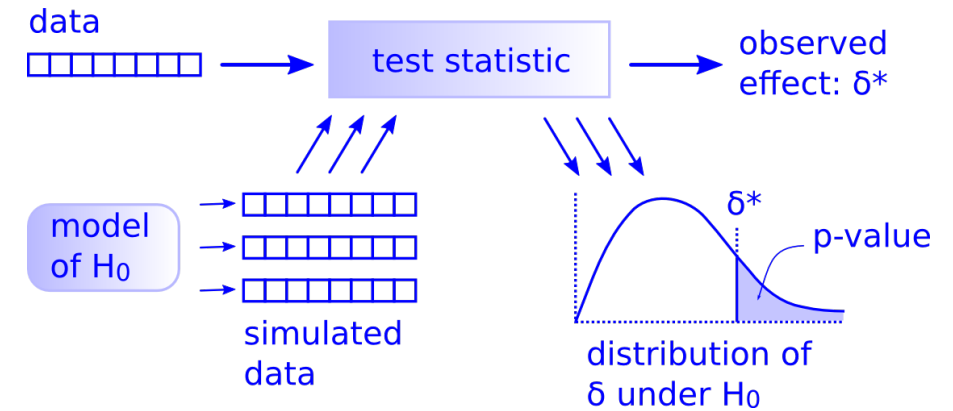
- Convention, p-value < 0.05
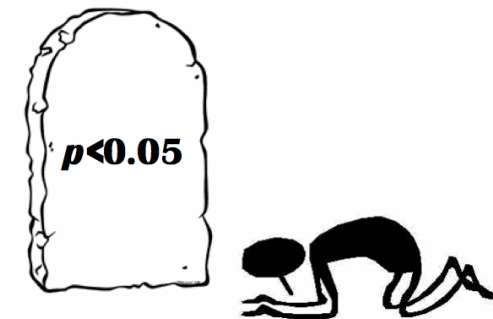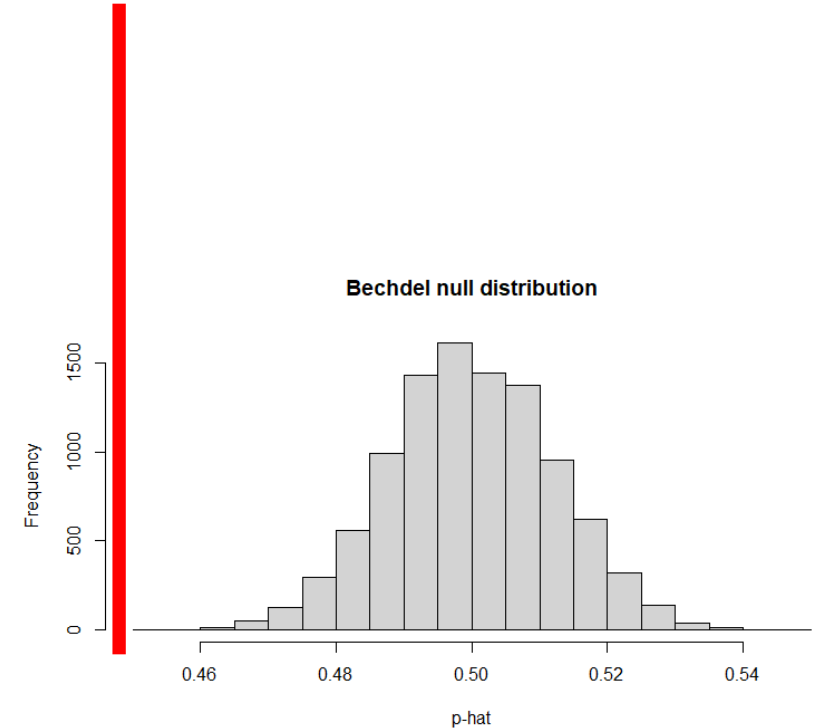
Bechdel null distribution

$p<0.05$

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

Bechdel (hypothesis) test

1. State the null hypothesis and the alternative hypothesis
   - 50% of the movies pass the Bechdel test: $H_0: \pi = 0.5$
   - Less than 50% of movies pass the: $H_A: \pi < 0.5$

2. Calculate the observed statistic
   - 803 out of 1794 movies passed the Bechdel test

$\hat{p} = .448$

3. Create a null distribution that is consistent with the null hypothesis
   - i.e., the statistics we expect if 50% of the movies passed the Bechdel test

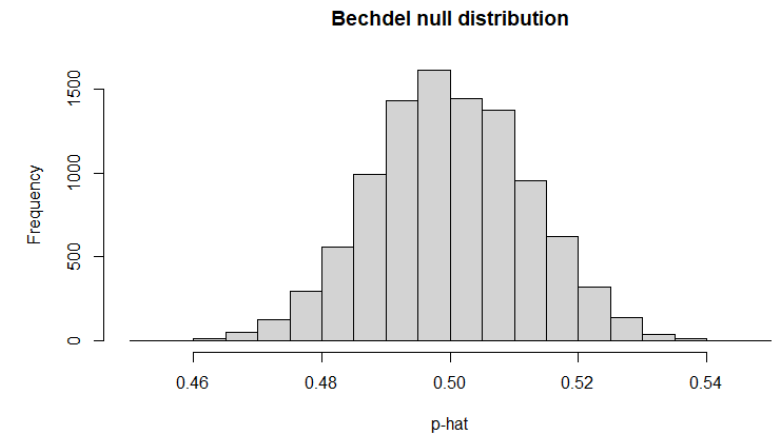4. Examine how likely the observed statistic is to come from the null distribution
   - What is the probability that only 803 of 1794 movies would pass the Bechdel test ($\hat{p} = .448$) if the null hypothesis was true?
   - i.e., what is the p-value?

5. Make a judgement
   - A small p-value this means that $\pi = .5$ is unlikely, and so it is likely $\pi < .5$
   - i.e., we say our results are 'statistically significant'

Hypothesis tests multiple proportions

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

The only difference is the parameters we are testing in step 1, and consequently the statistics we use...
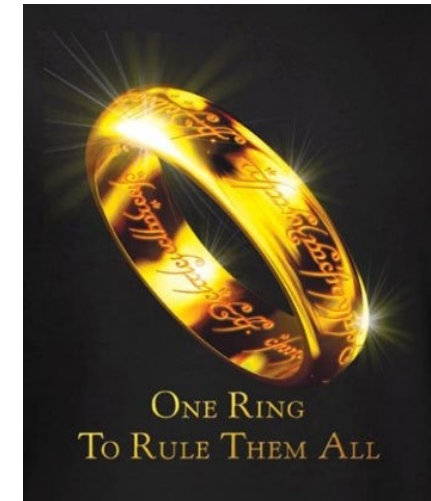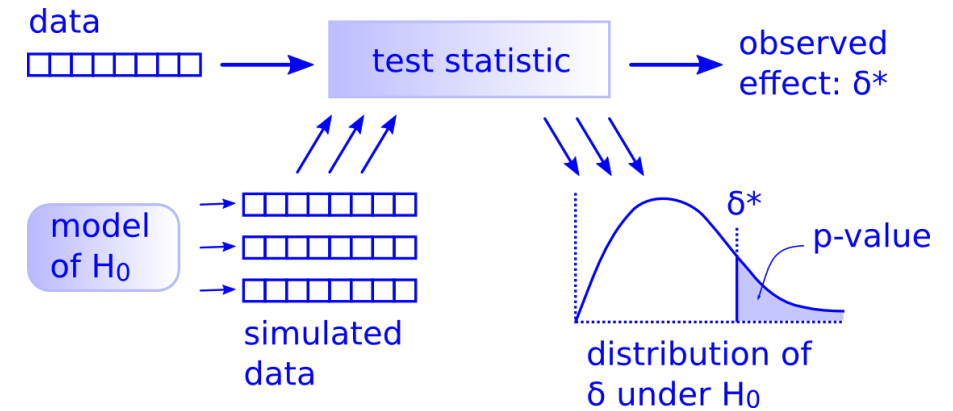
Example: Jury selection in Alameda county

Section 197 of California's Code of Civil Procedure says:

" All persons selected for jury service shall be selected at random, from a source or sources inclusive of a representative cross section of the population of the area served by the court."

Eligible jurors in a County ▶ List of eligible residents ▶ **Jury panel** ▶ Jury
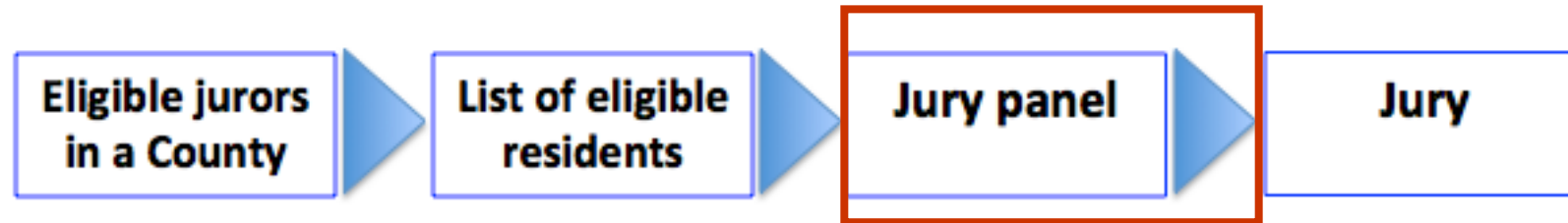
In 2010, the American Civil Liberties Union (ACLU) of Northern California presented a report that concluded that certain racial and ethnic groups are underrepresented among jury panelists in Alameda County.

RACIAL AND ETHNIC DISPARITIES IN ALAMEDA COUNTY JURY POOLS

A Report by the ACLU of Northern California   October 2010

Step 1: Null and Alternative hypothesis

The null hypothesis is that the proportion of people on jury panels matches the underlying demographics

We can write the null hypothesis in symbols using:

- $\pi_{\text{Asian-on-panels}} = .15$
- $\pi_{\text{Latino-on-panels}} = .12$
- etc.

Proportions in the population

The alternative hypothesis that the proportion of at least one ethnicity does not match the underlying population

We can write this using symbols as: at least one π_i is not as specified

Step 2: Calculating the observed statistic

The ACLU compiled data on the composition of 1453 people who were on jury panels from in the years 2009 and 2010.

People on the panels are of multiple ethnicities
- Distribution of ethnicities is categorical

To see whether the distribution of ethnicities of the panels is close to that of the eligible jurors, we have to measure the distance between two categorical distributions
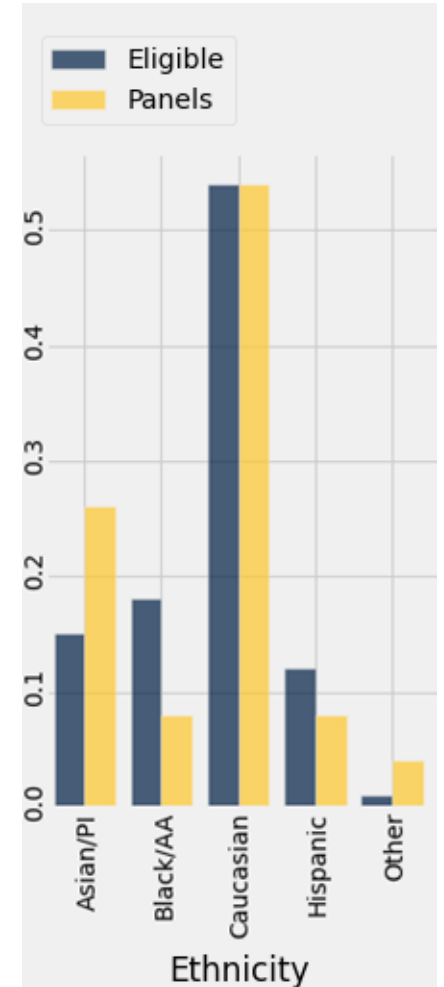
Total variation distance

To run a hypothesis test we need to select a statistic

A statistic we can use to measure the deviation of two distributions of proportions is the **Total Variation Distance** (TVD) which can be calculated using:

- For each category, compute the difference in proportions between two distributions

- Take the absolute value of each difference

- Sum the values

$$TVD = \sum_{i=1}^{k} |\pi_i - \hat{p}_i|$$

The value of the TVD statistic for Almeda county is 0.28
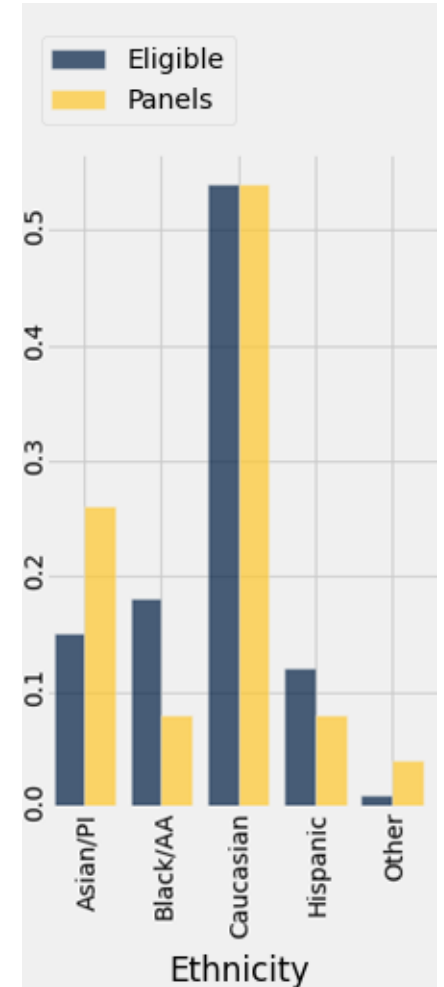
Step 3: Creating a null distribution

To create a null distribution, we need to randomly generate several proportions consistent with the null hypothesis
- i.e., $\hat{p}_{Asian}$, $\hat{p}_{Latino}$ etc.

We can do this by randomly generating numbers between 0 and 1, and then splitting the data at the cumulative sums of the proportions specified by the null hypothesis
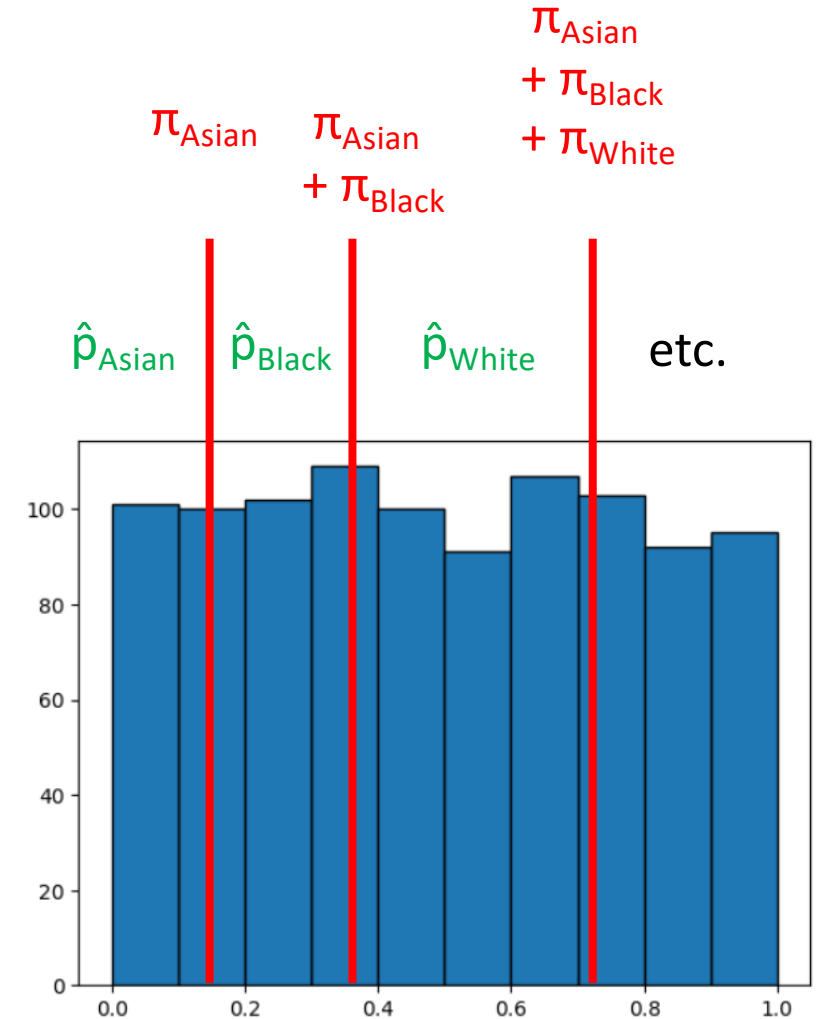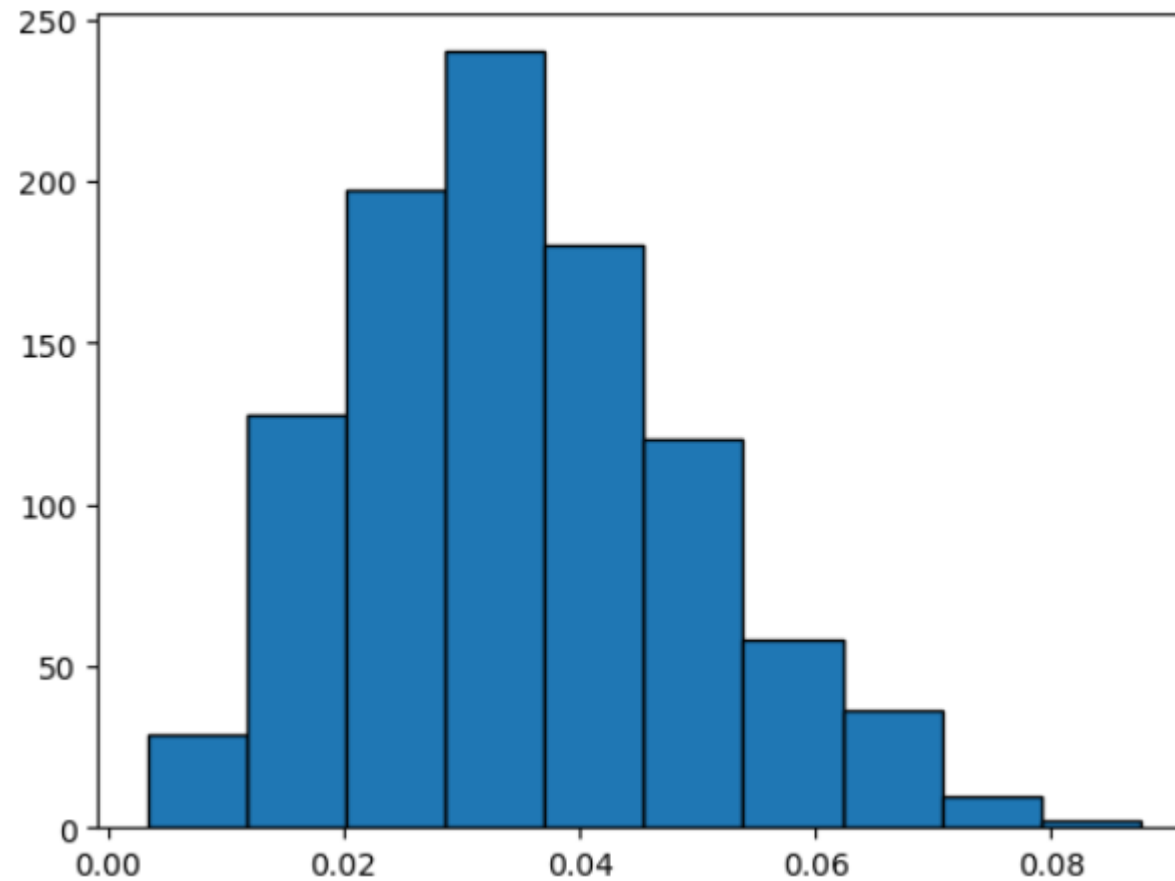
Step 3: Creating a null distribution

One we have generated $\hat{p}_{Asian}$, $\hat{p}_{Latino}$ etc. consistent with the null hypothesis, we can then calculate the TVD between these random and the true $\hat{p}$'s and the π_i's specified by the null hypothesis

$$TVD = \sum_{i=1}^{k} |\pi_i - \hat{p}_i|$$

We can repeat this 10,000 times to get a null distribution...

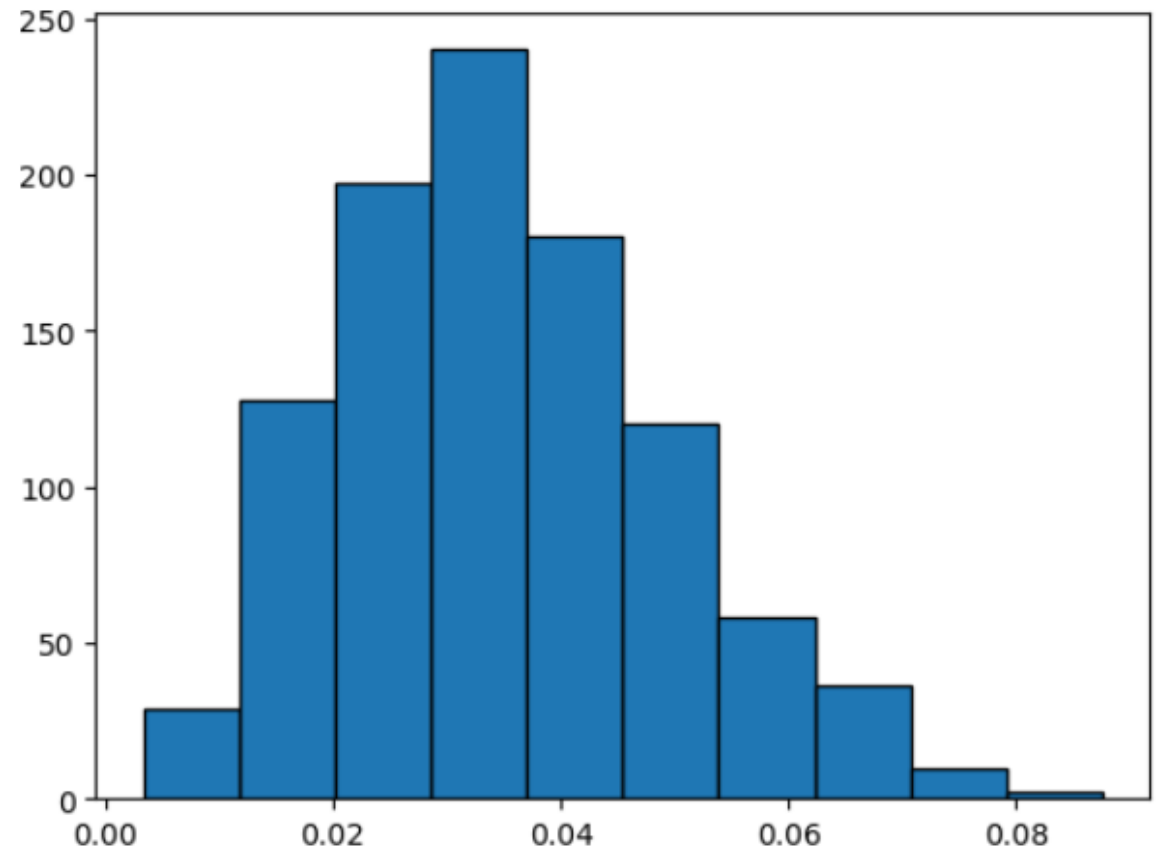# Step 3: Creating a null distribution

Step 4: Calculate the p-value

The p-value is the proportion of statistics in the null distribution more extreme than our observed statistic

Our observed statistic TVD value was 0.28

What is the p-value?

Step 5: Draw a conclusion

A small p-value is evidence to reject the null hypothesis
- i.e., our data is not consistent with the null hypothesis

Thus, we can conclude that the ethnicities of members on jury panels do not accurately reflect the underlying demographics.

Potential reasons for bias in Alameda county jury selection

Rejection of model tells us the model doesn't accurately account for the data, but it doesn't tell us why

The ACLU identified several reasons for bias in jury selection including:

- The software didn't work well, contributing to biased selection
- Jurors were selected at random from everyone who is a registered voter and/or has a driver's license
- Hard to reach people who don't have permanent addresses
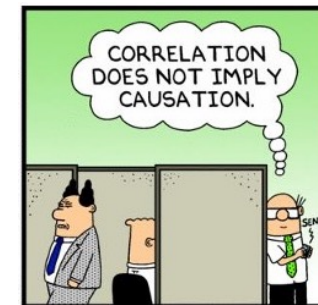 - Can disproportionately affect people at lower income levels
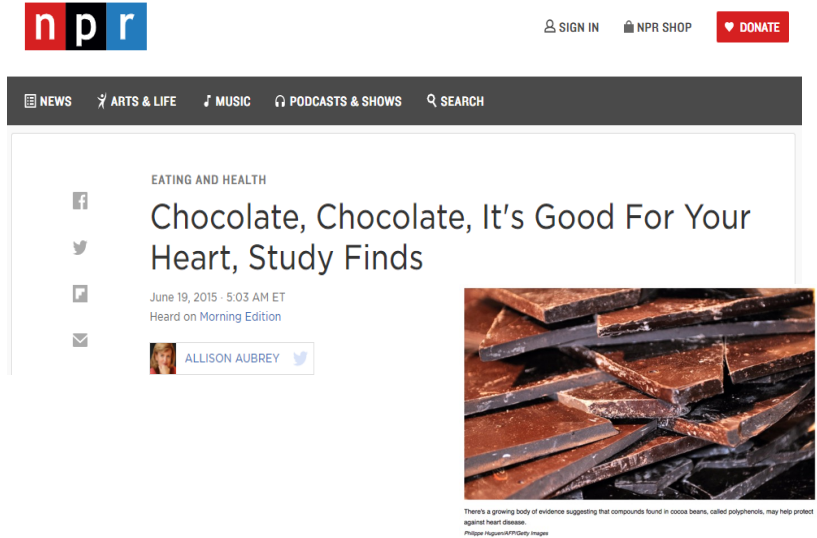
Let's explore this in Jupyter!

Assessing causal relationships

Review: Causality

Recall from class 2:

- **An association** is the presence of <u>a reliable relationship</u> between the treatments an outcome

- **A causal relationship** is when changing the value of a treatment variable <u>influences</u> 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?

CORRELATION DOES NOT IMPLY CAUSATION.

Lurking variable

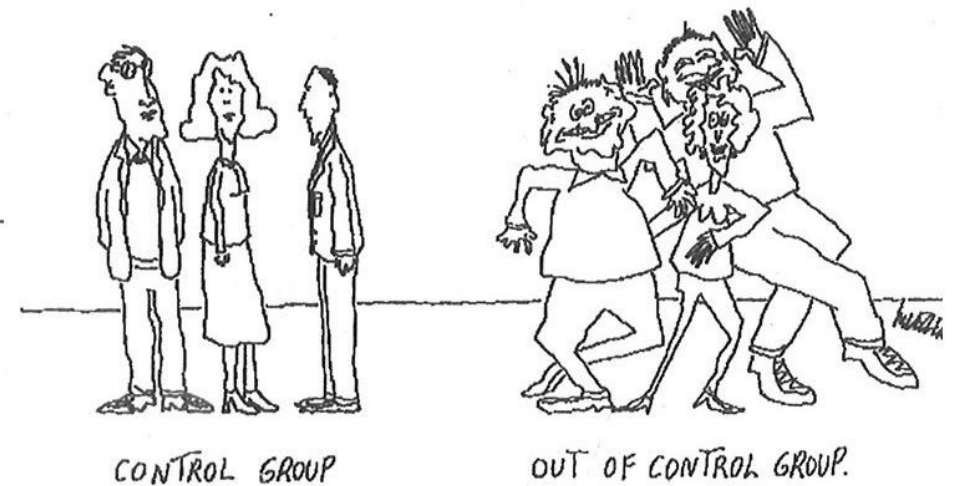# Randomized Controlled Experiment

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

CONTROL GROUP   OUT OF CONTROL GROUP.

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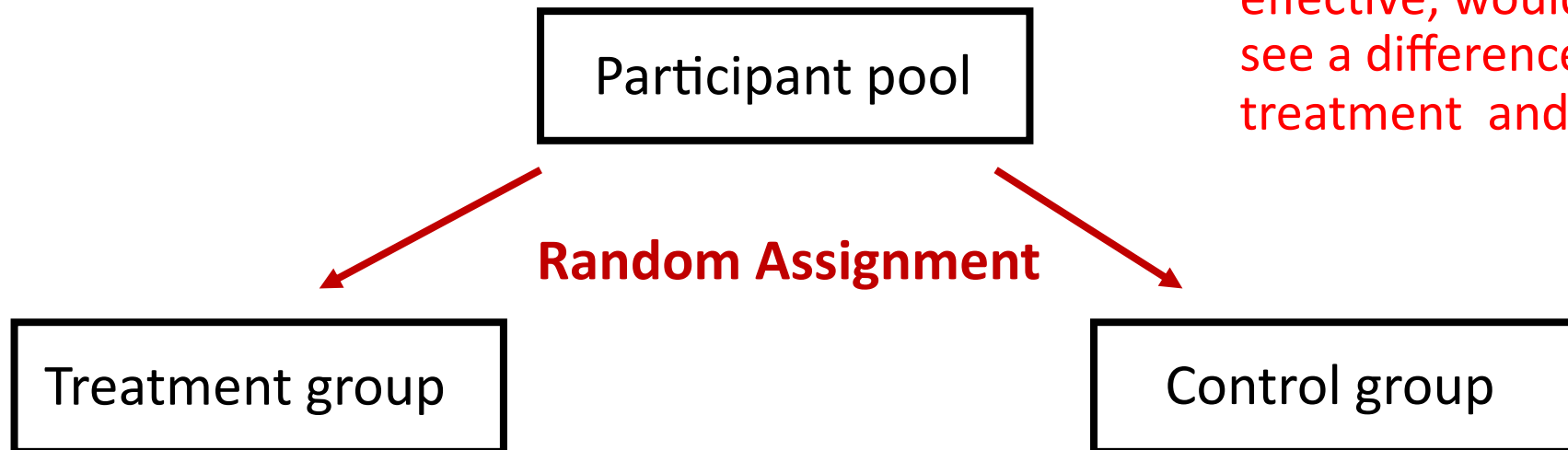
Q: If the treatment was not effective, would we expect to see a difference between the treatment and control groups?

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?

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: $\pi_{treat} > \pi_{control}$

Step 2: The observed statistic

To calculate an observed statistic we need data:
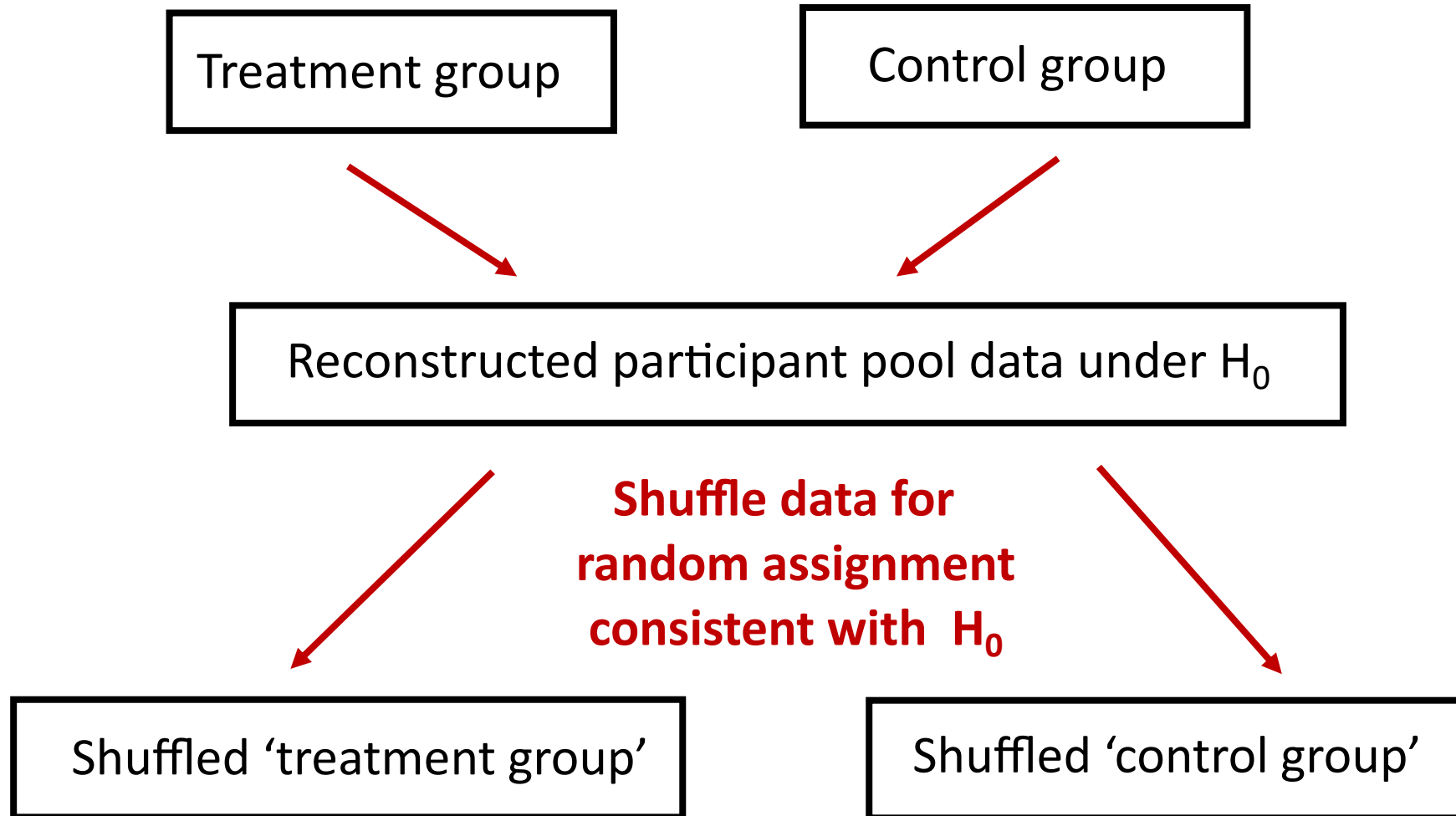
Let's have our observed statistic mirror our hypotheses

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

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
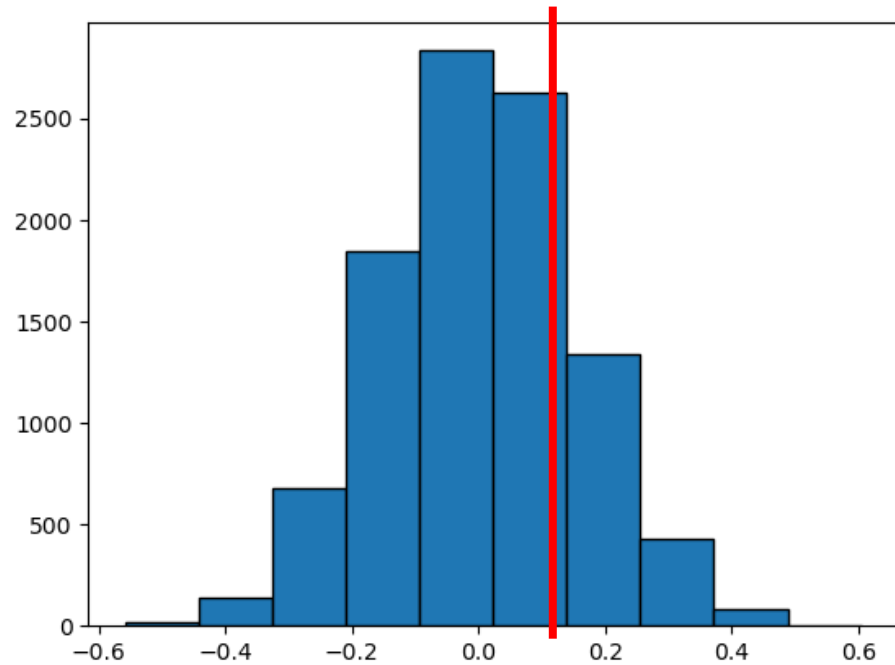
Shuffled 'treatment group'

Shuffled 'control group'

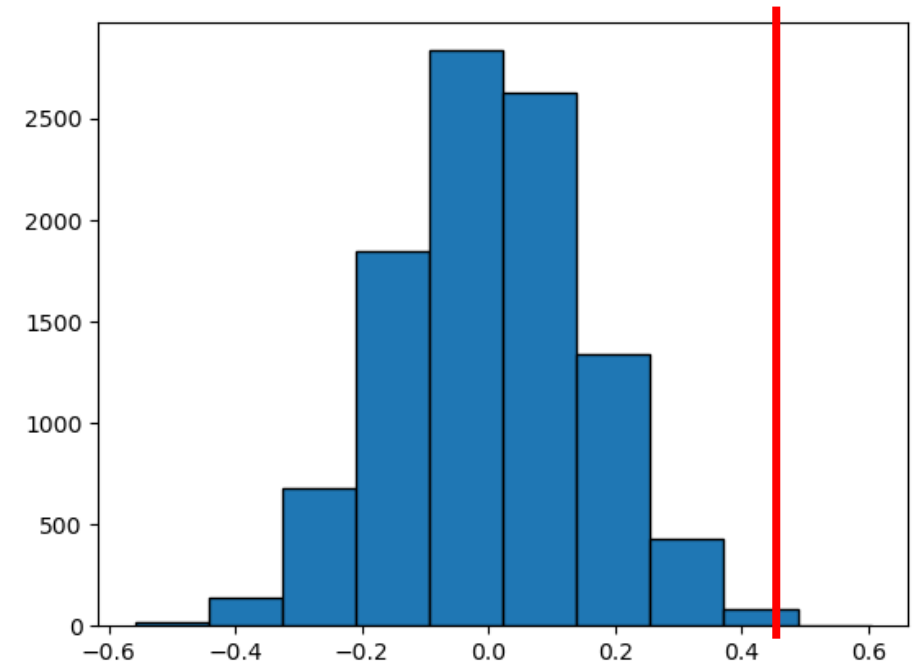One null distribution statistic: $\hat{p}_{Shuff\_Treatment} - \hat{p}_{Shuff\_control}$

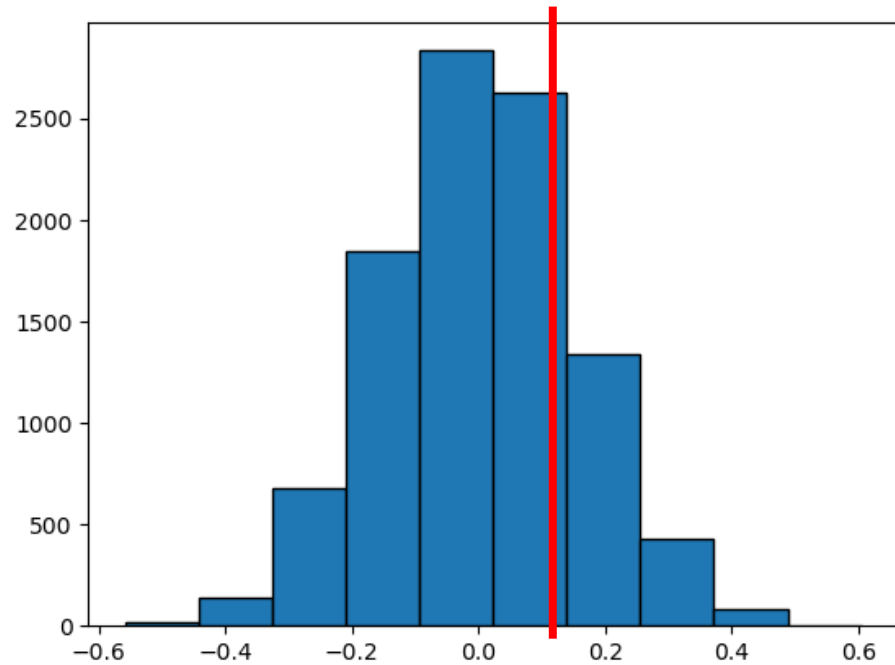# Step 3: Create a null distribution
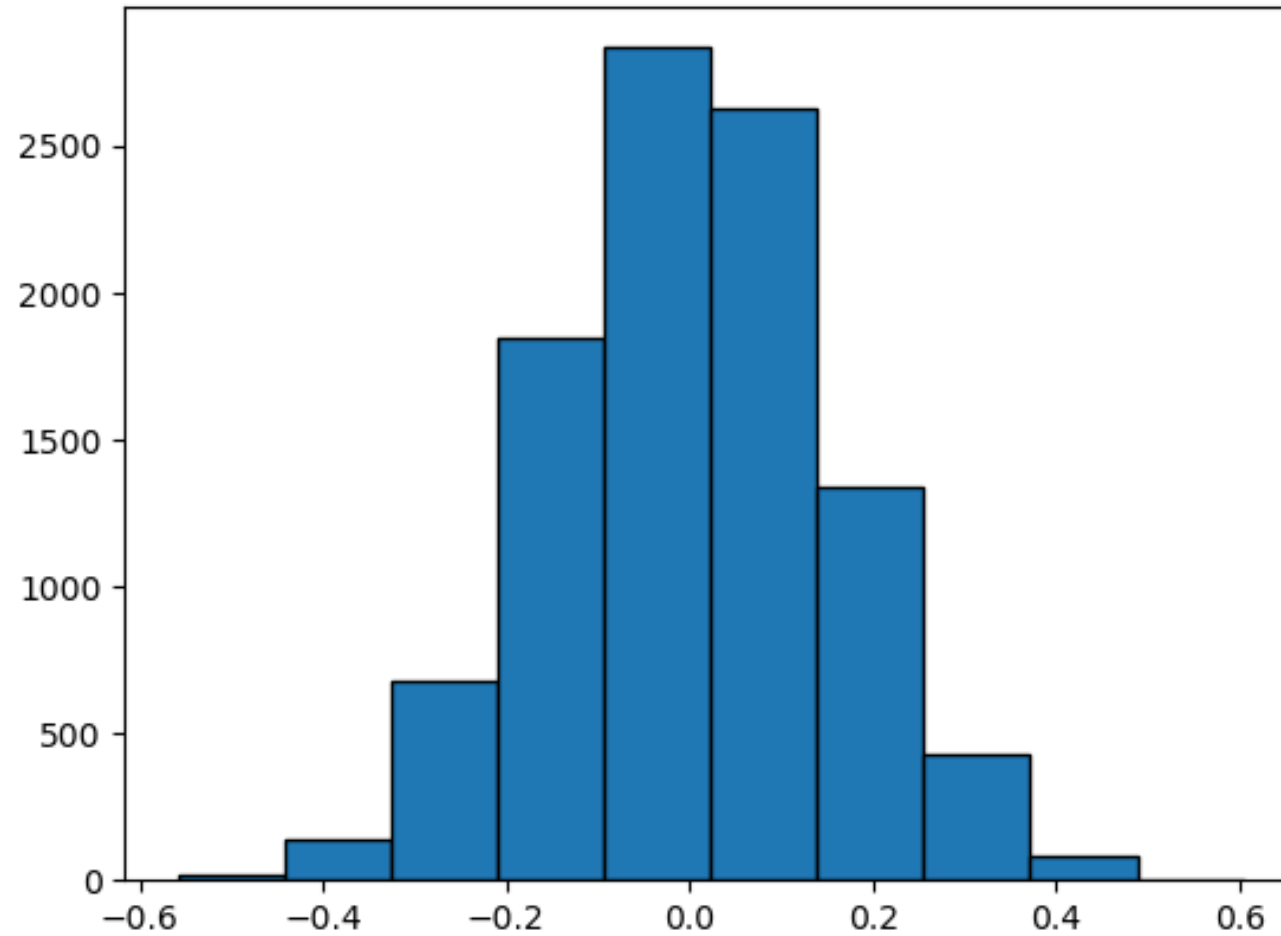
Step 4: Calculate the p-value

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

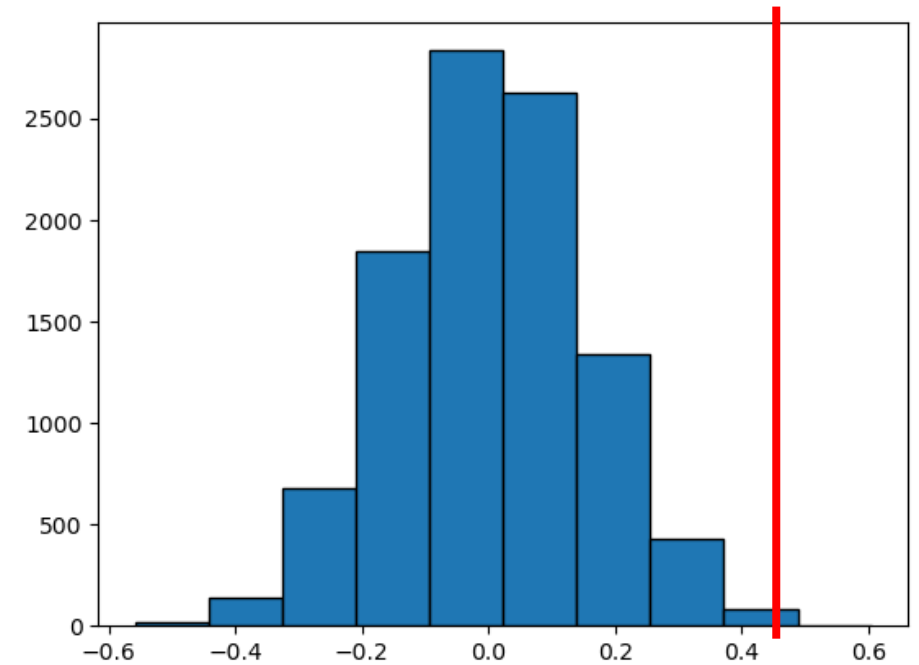If $\hat{p}_{treat} - \hat{p}_{control} = 0.5$ what would the p-value be?

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?

Let's explore this in Jupyter!