

# YData: Introduction to Data Science



Class 20: Hypothesis tests

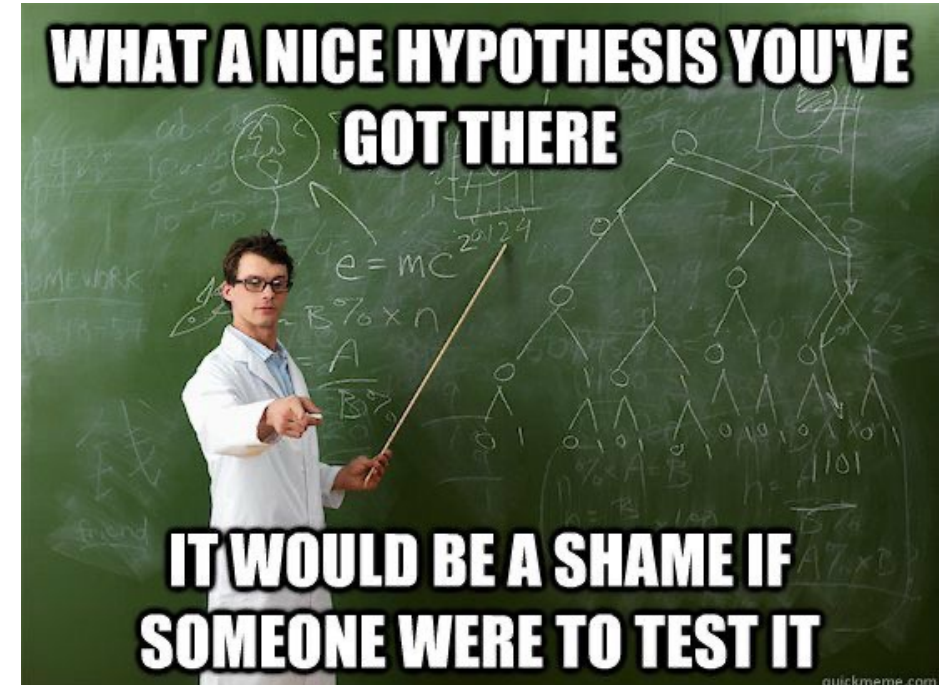
# Overview

Project related topic: dealing with dates

Review of parameters, statistics, and sampling

Hypothesis tests for a single proportion  
continued

If there is time: hypothesis tests for multiple  
proportions



# Project timeline

## Tuesday, April 11<sup>th</sup>

- 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 will be posted to Canvas

## Wednesday, April 19<sup>th</sup>

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

## Sunday, April 30<sup>th</sup>

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



# Dealing with dates

Often there are columns in a DataFrame that contain information about dates

We can read date columns as “datetime” types into a DataFrame by setting the `parse_date` argument in `read_csv()`.

```
my_data = pd.read_csv("file.csv", parse_dates = [0])    # column 0 has dates
```

We can also create datetime objects using the datetime module:

```
import datetime
my_date = datetime.datetime(2000, 1, 1)
dow2.iloc[:, 0] < my_date    # get all dates before my_date
```

Let's explore this in Jupyter!

# Review of Statistical Inference

# Inference

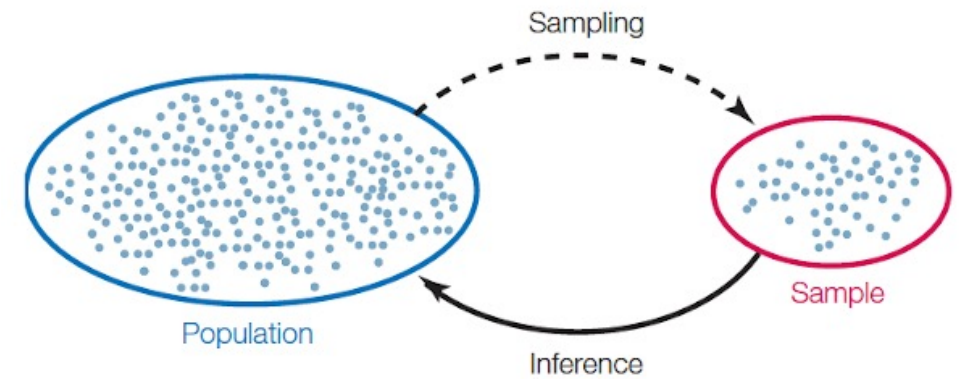
**Statistical Inference:** Making conclusions about a population based on data in a random sample

This usually involves using data in a sample to estimate the value of a **fixed** unknown number

- i.e., we estimate values of a “parameter”

Example:

- Estimating the average height of all humans on Earth from a random sample of 1,000 humans
  - Our estimate will vary from sample to sample



# Terminology

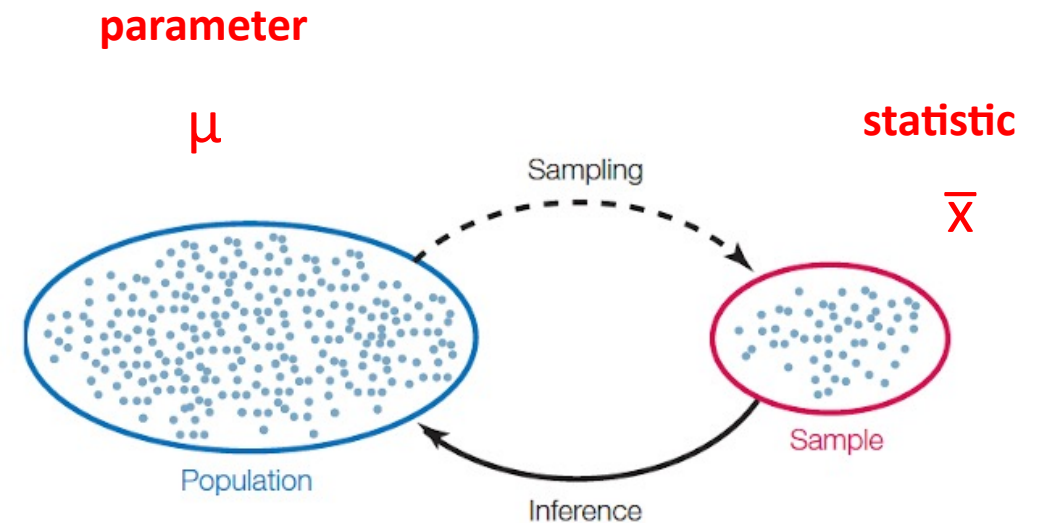
A **parameter** is number associated with the population

- e.g., population mean  $\mu$
- e.g., average height of all humans

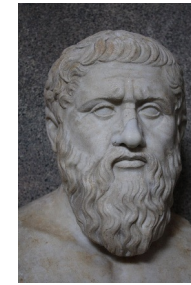
A **statistic** is number calculated from the sample

- e.g., sample mean  $\bar{x}$
- e.g., average height of 1,000 people in our sample

A statistic can be used as an estimate of a parameter



# Examples of parameters and statistics



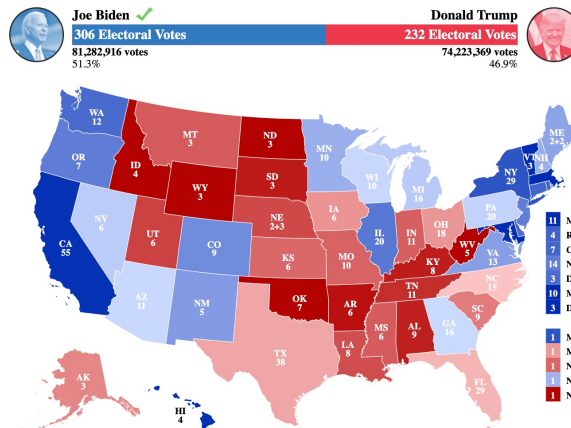
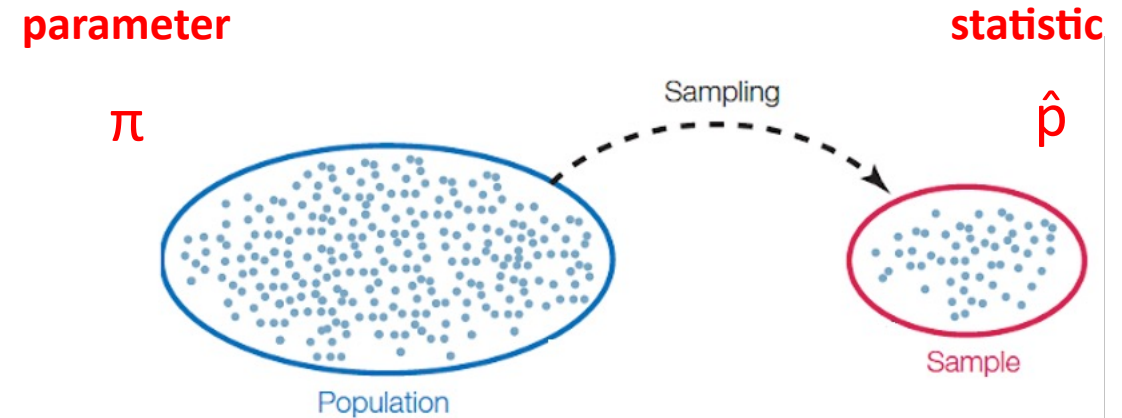
|                    | Sample Statistic | Population Parameter |
|--------------------|------------------|----------------------|
| Mean               | $\bar{x}$        | $\mu$                |
| Standard deviation | $s$              | $\sigma$             |
| Proportion         | $\hat{p}$        | $\pi$                |
| Correlation        | $r$              | $\rho$               |
| regression slope   | $b$              | $\beta$              |



# Sampling

**Simple random sample:** each member in the population is equally likely to be in the sample

- Allows for generalizations to the population



Polls of 1,000 voters:  $\hat{p}_{\text{Biden}}$

Vote on election day:  $\pi_{\text{Biden}}$

# Probability distribution of a statistic

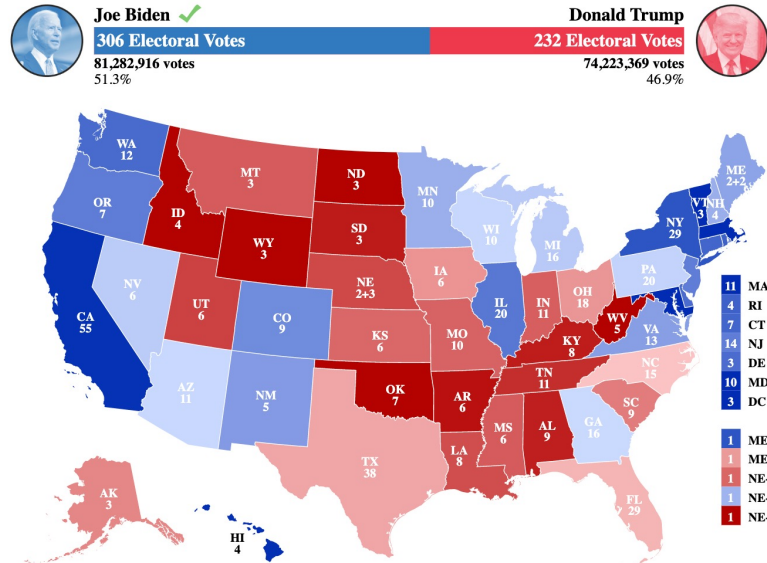
Values of a statistic vary because random samples vary

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

$\pi_{\text{Biden}}$

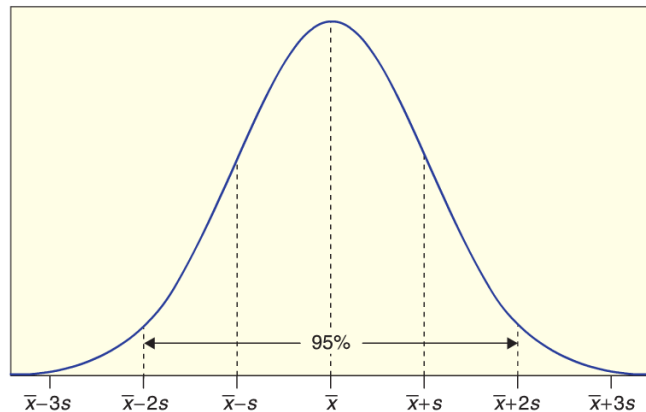
$n = 1,000$



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



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



Sampling distribution!



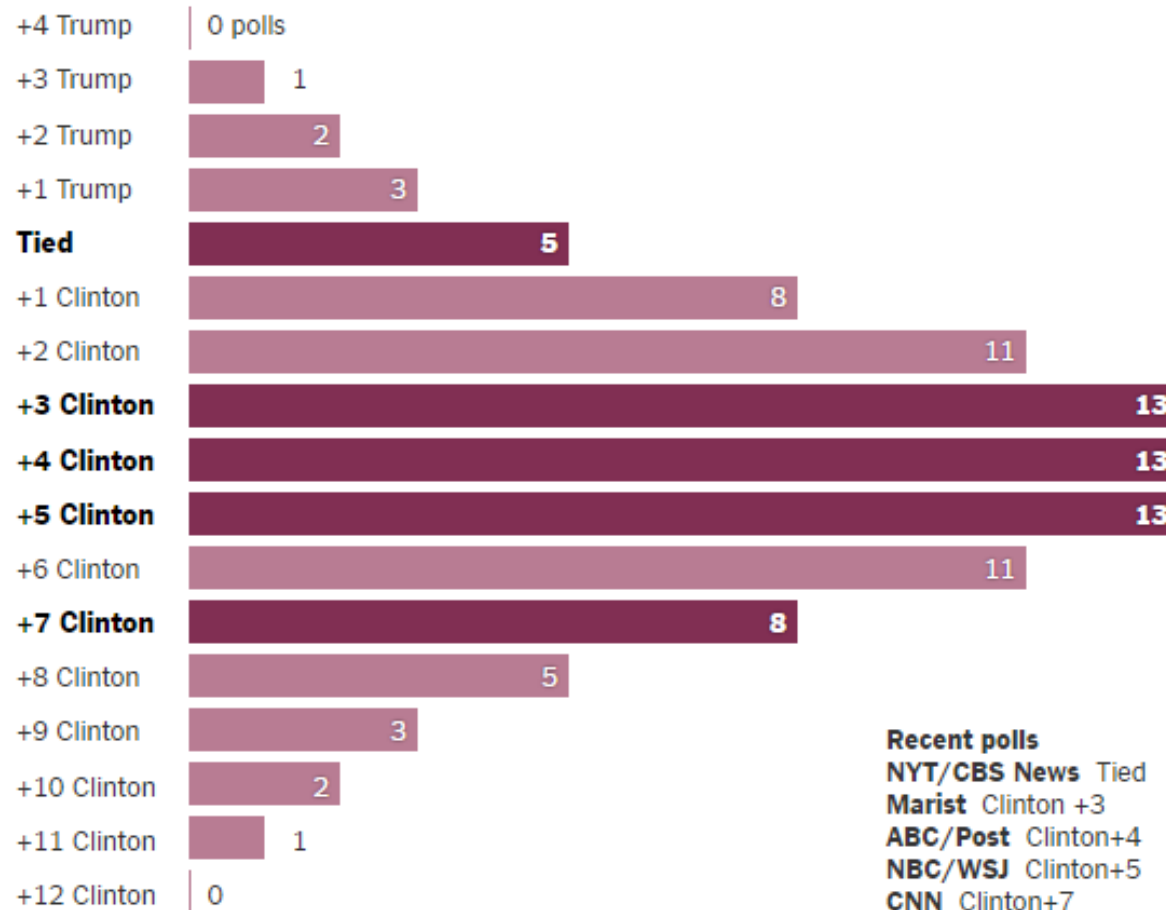
$\hat{p}_{\text{Biden}}$

# Confused by Contradictory Polls? Take a Step Back

## Noisy Polls Are to Be Expected

If Hillary Clinton were up by a modest margin, there would be plenty of polls showing a very close race — or even a Trump lead.

**A simulation of 100 surveys, if Mrs. Clinton were really up 4 points nationally.**



What is this called?



What parameter are they trying to estimate?

# Simulating random proportions ( $\hat{p}$ 's)

We can simulate random proportions  $\hat{p}$  consistent with a population proportion  $\pi$  by:

1. Generated  $n$  random numbers uniformly distributed between 0 and 1
  - `rand_nums = np.random.rand(1000)`
2. Marking points less than  $\pi$  as being **True**, and greater  $\pi$  than as being **False**
  - `rand_binary = rand_nums <= pi_value`
3. Calculating the proportion of points to get a  $\hat{p}$ 
  - `rand_phat = np.mean(rand_binary)`



Let's explore this in Jupyter!

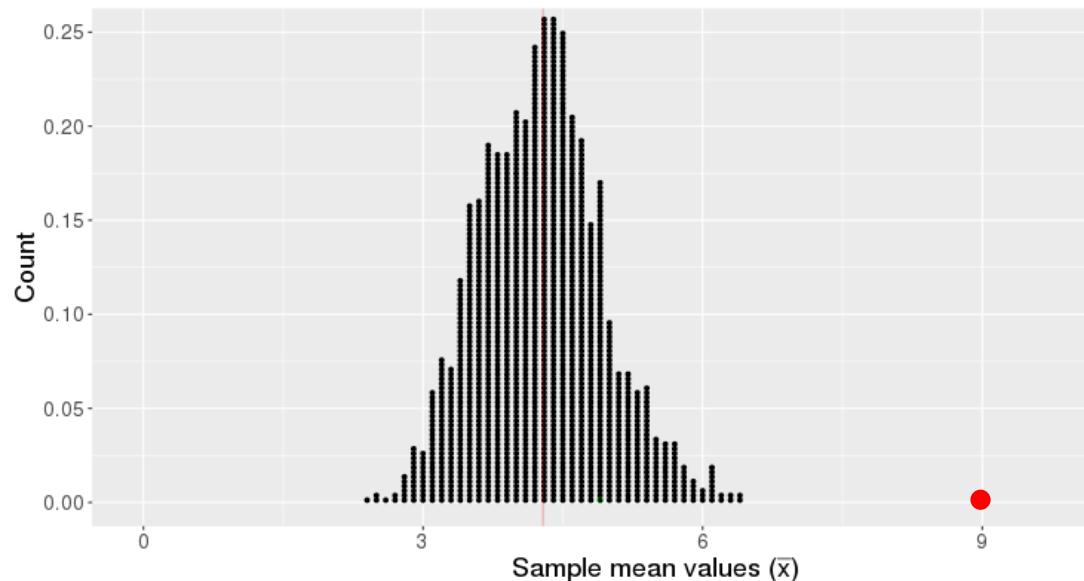
# Hypothesis tests

# 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

# Example claims (hypotheses)

Let's see if we can write the following claims (hypotheses) using symbols

Claim: 88% of Yale students graduate within four years

- $H: \pi = 0.88$

Claim: The average age of a Yale undergraduate is 20

- $H: \mu = 20$

Claim: 70.7% of Yale classrooms have fewer than 20 students in attendance

- $H: \pi = 0.707$



# Testing claims (hypotheses)

Claim: 88% of Yale students graduate within four years

- $H: \pi = 0.88$
- To test this claim, we could randomly selected  $n = 100$  Yale graduates.
- If we found the proportion that graduated in 4 years is  $\hat{p} = .80$ , would we believe the claim?

# Testing claims (hypotheses)

Claim: The average age of a Yale undergraduate is 20

- $H: \mu = 20$
- To test this claim, we could randomly selected  $n = 50$  Yale graduates.
- If we found the average age of in our sample of students was  $\bar{x} = 20.2$ , would we believe the claim?

# Motivating example: The Bechdel Test



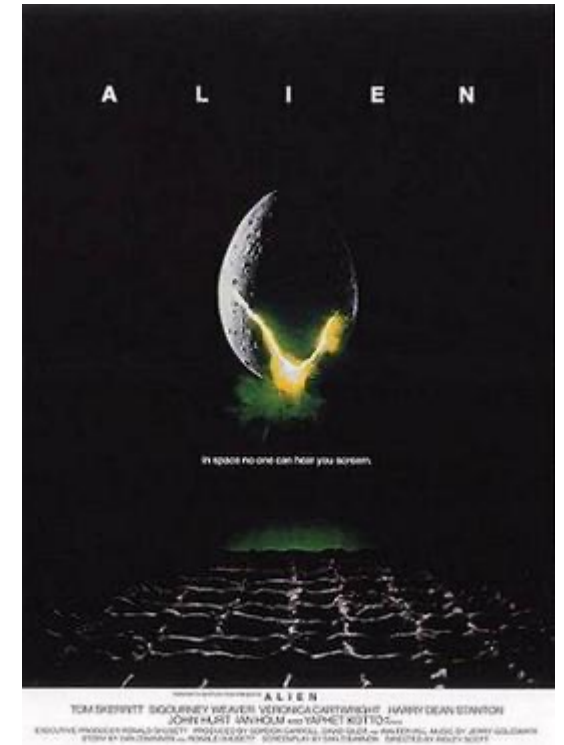
For a movie to pass the Bechdel Test it must meet three criteria:

1. It has to have at least 2 women in it
2. The women must talk to each other
3. They must talk about something besides a man

# Motivating example: The Bechdel Test

Suppose we had a random sample of 1794 movies

- The *sample size* is 1794 ( $n = 1794$ )



# Motivating example: The Bechdel Test

**Question:** Do less than 50% of movies pass the Bechdel test?

## Questions:

- What is the population/process?
- What is our parameter of interest?
  - What symbol should we use to denote it?
- What is our statistic of interest?
  - What symbol should we use to denote it?

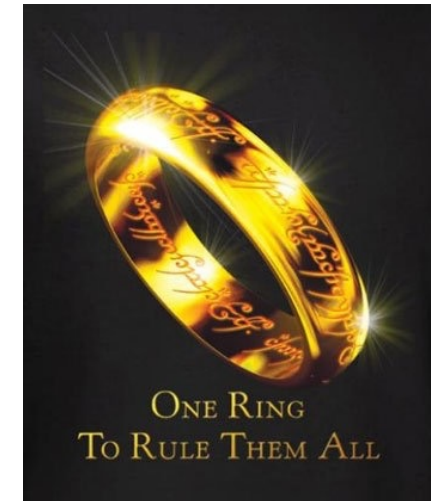
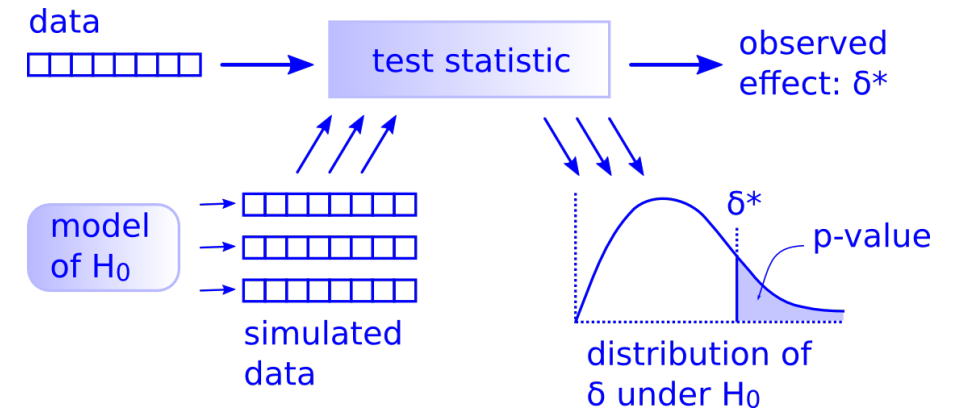
|    | title                  | binary |
|----|------------------------|--------|
| 1  | Dredd 3D               | PASS   |
| 2  | 12 Years a Slave       | FAIL   |
| 3  | 2 Guns                 | FAIL   |
| 4  | 42                     | FAIL   |
| 5  | 47 Ronin               | FAIL   |
| 6  | A Good Day to Die Hard | FAIL   |
| 7  | About Time             | PASS   |
| 8  | Admission              | PASS   |
| 9  | After Earth            | FAIL   |
| 10 | American Hustle        | PASS   |
| 11 | August: Osage County   | PASS   |
| 12 | Beautiful Creatures    | PASS   |
| 13 | Blue Jasmine           | PASS   |
| 14 | Captain Phillips       | FAIL   |

# 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

Let's go through these steps now...



# Do less than 50% of movies pass the Bechdel test?

Step 1: state the null and alternative hypotheses

If 50% of the movies passed the Bechdel test, what would we expect the value of the parameter to be?

$$H_0: \pi = 0.5$$

If fewer than 50% of movies passed the Bechdel test, what would we expect the value of the parameter to be?

$$H_A: \pi < 0.5$$

# Step 2: Calculate the observed statistic value

Step 2: calculate the observed statistic

There are 1794 movies in our data set

Of these, 803 passed the Bechdel test

What is our observed statistic value and what symbol should we use to denote this value?

A:  $\hat{p} = 803/1794 = 0.448$



# Step 3: Create a null distribution

How can we assess whether 803 out of 1794 movies passing the Bechdel test ( $\hat{p} = 0.448$ ) is consistent with what we would expect if 50% (or more) movies passed the Bechdel test?

- i.e., is  $\hat{p} = 0.448$  a likely value if  $\pi = 0.5$ ?

If 50% of movies passed the Bechdel test, we can model movies passing the as a fair coin flip:

Heads (True) = passed the Bechdel test

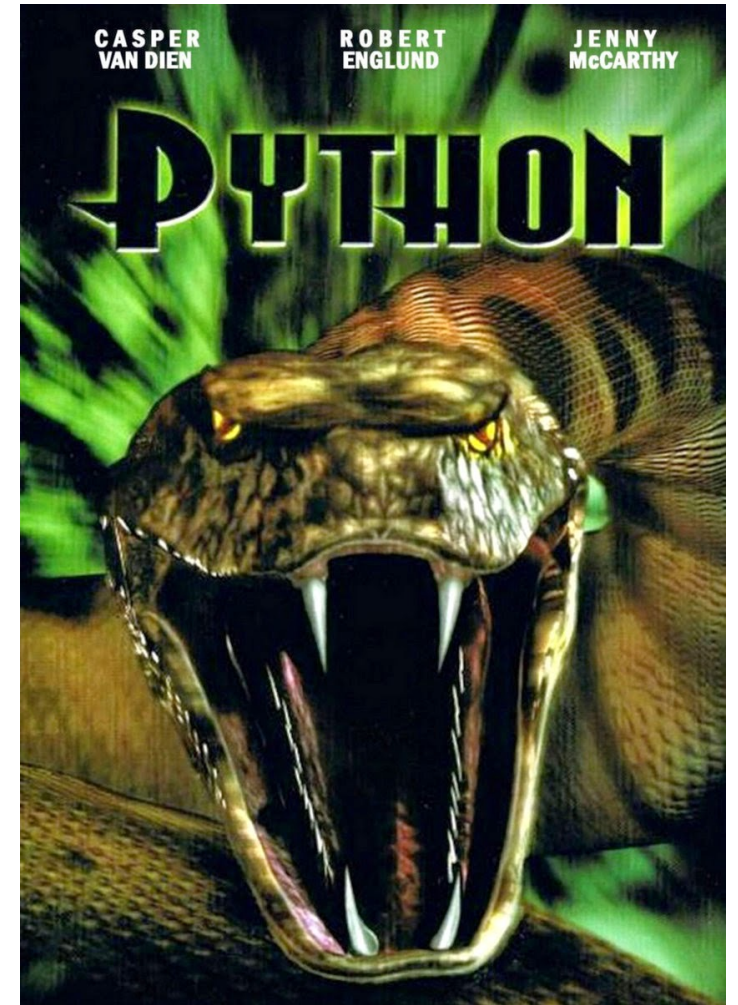
Tails (False) = failed to pass the Bechdel test

Let's simulate flipping a coin 1794 times and see how many times we get 803 **or fewer** heads

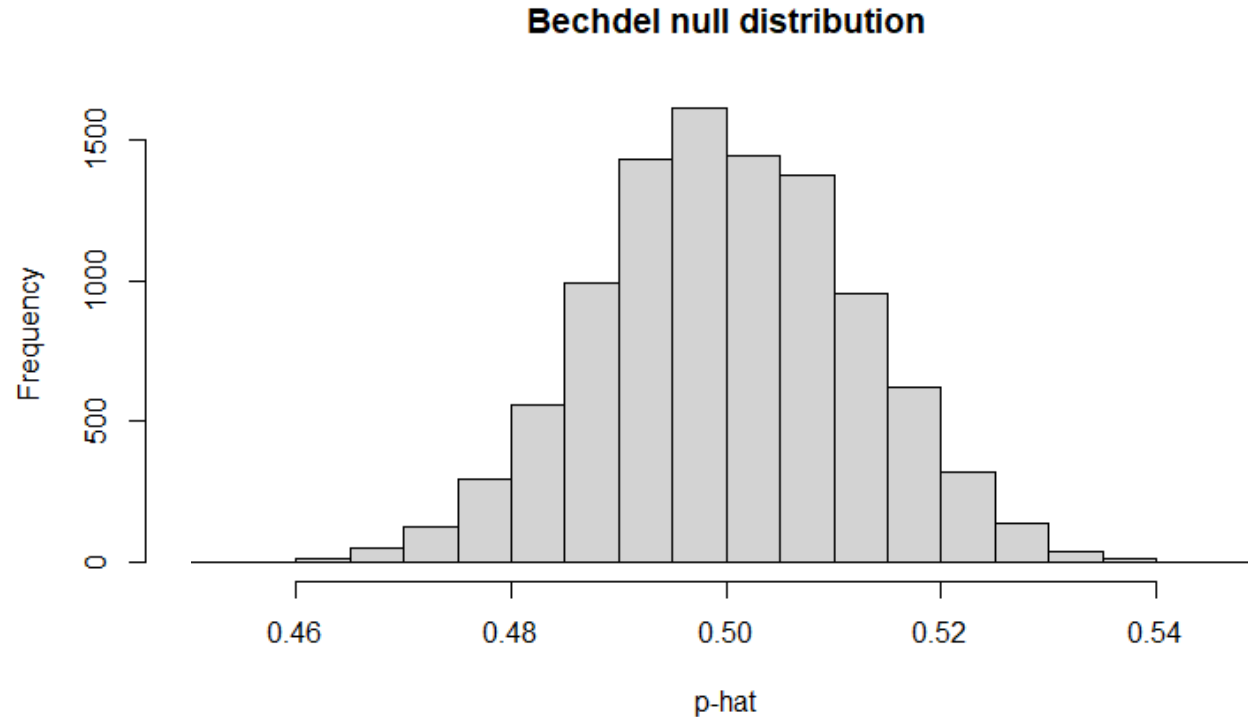
# Chance models

To really be sure, how many repetitions of flipping a coin 1794 times should we do?

Any ideas how to do this?

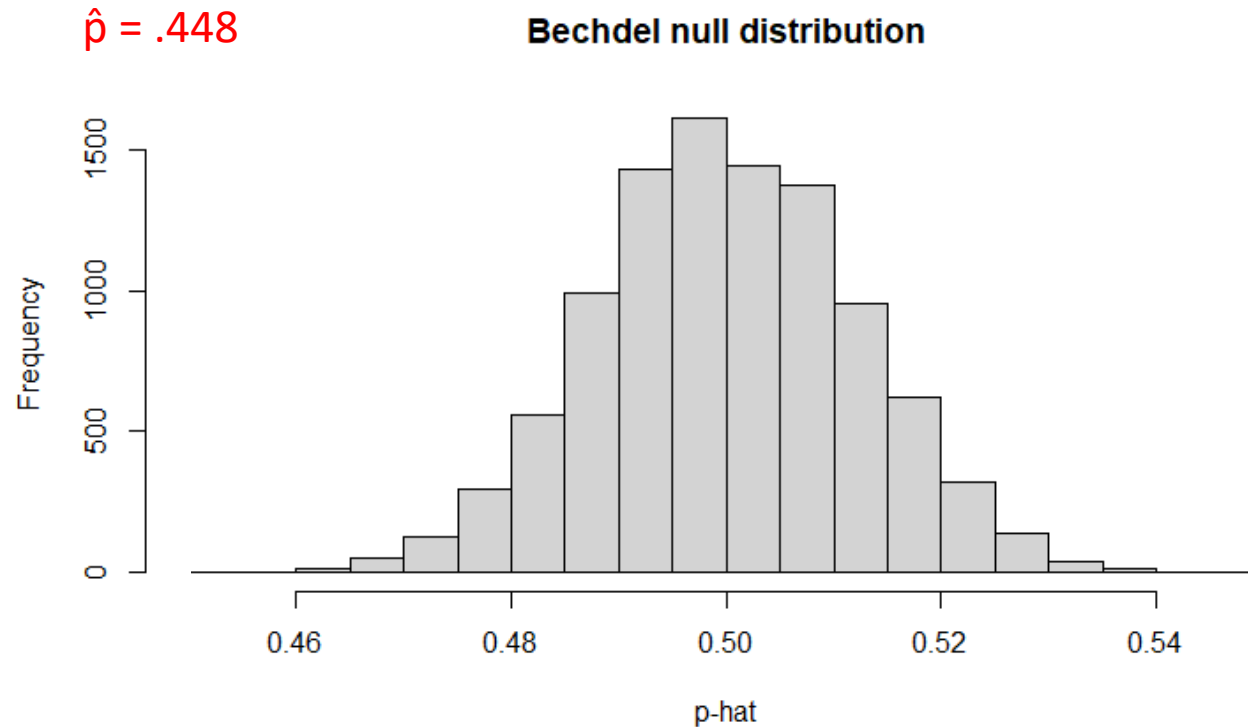


# Simulating Flipping 1794 coins 10,000 times



Assuming the null hypothesis is true, the distribution of statistics we get is called the **null distribution**

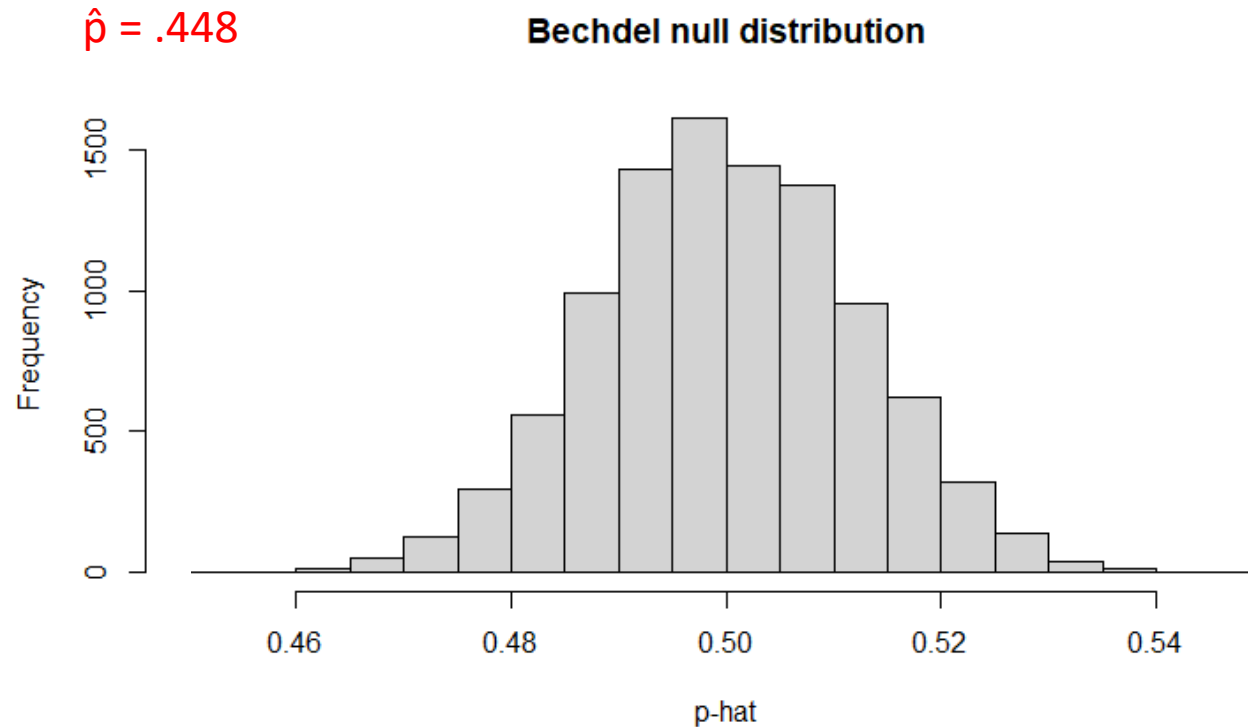
# Step 4: calculate the p-value



Q: Is it likely that 50% of movies pass the Bechdel test?

- i.e., is it likely that  $\pi = .5$ ?

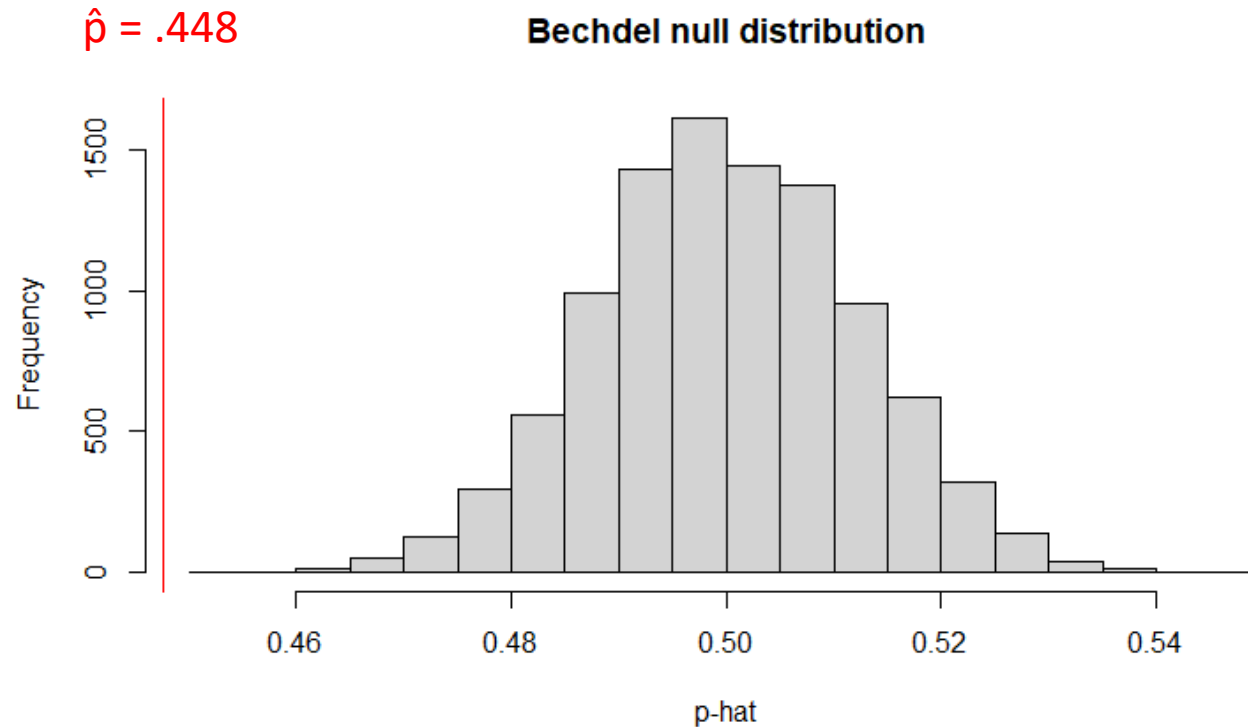
# Step 4: calculate the p-value



The **p-value** is the probability we will get a statistic as or more extreme than the observed statistic, if the null hypothesis was true

Q: What is the p-value here?      A: the p-value is 0

# Step 5: Make a decision



If the observed data is very unlikely if the null hypothesis is true, we can reject the null hypothesis

- i.e., if p-value is very small we can reject the null hypothesis

Let's try it in Python



# Key steps hypothesis testing

## 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

- Buzz got 803 out of 1794 movies passed the Bechdel test, or  $\hat{p} = .448$

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

- i.e., what statistics would 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

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



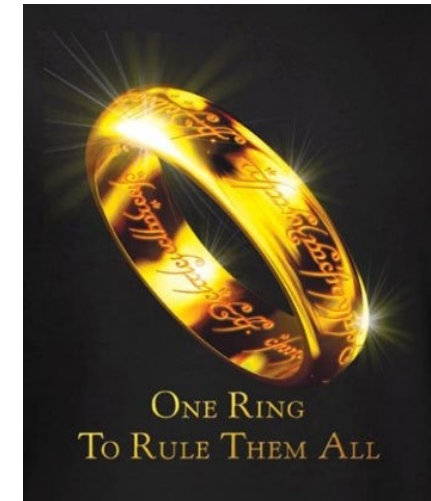
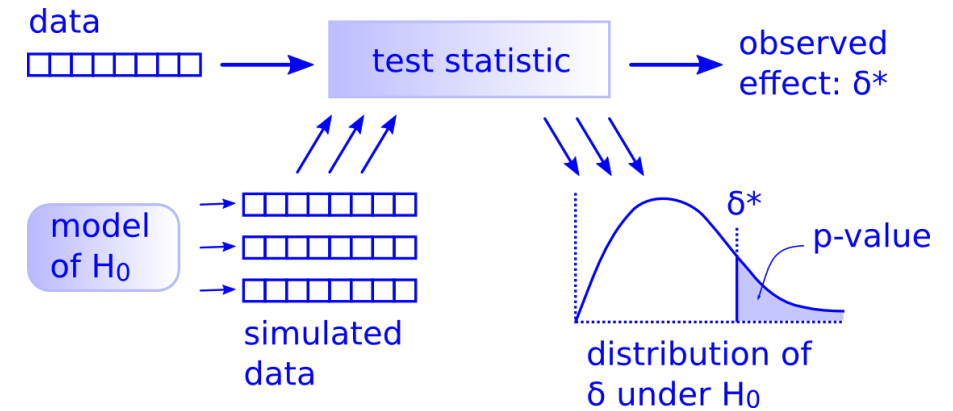
Hypothesis tests multiple proportions

# 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

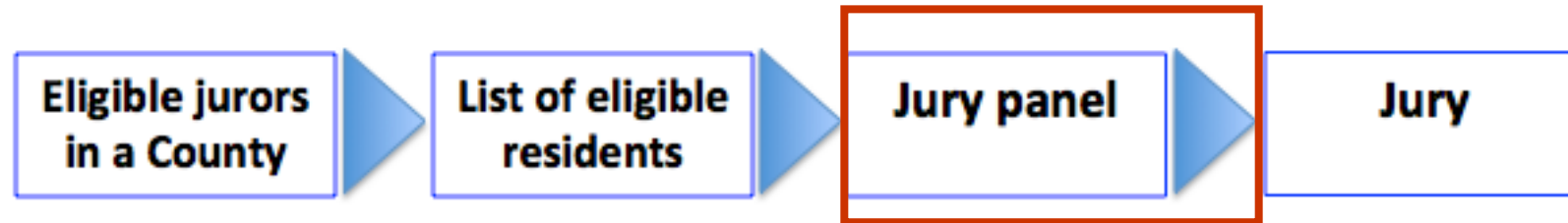
The only difference is the parameters we are testing in step 1, and consequently the statistics we use...



# 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."




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**

# 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:

- $\pi_{\text{Asian-on-panels}} = .15$
  - $\pi_{\text{Black-on-panels}} = .18$
  - 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  $\pi_i$  is not as specified

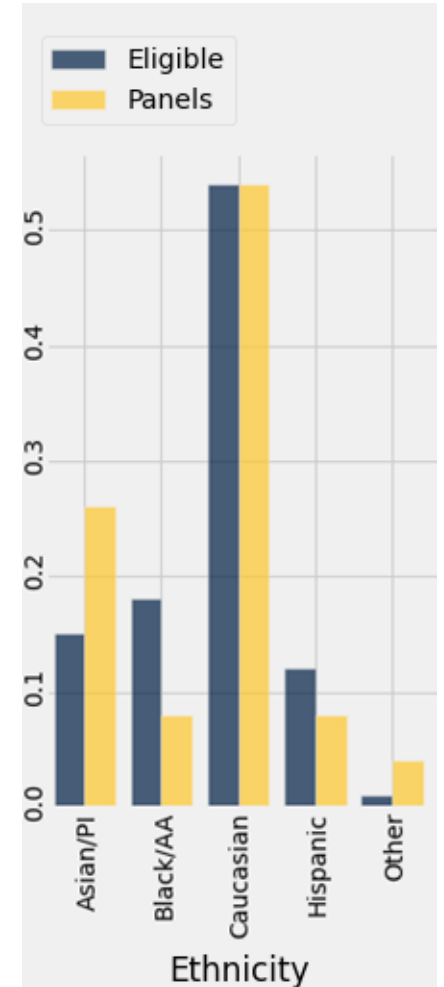
## Step 2: Calculating the observed statistic

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

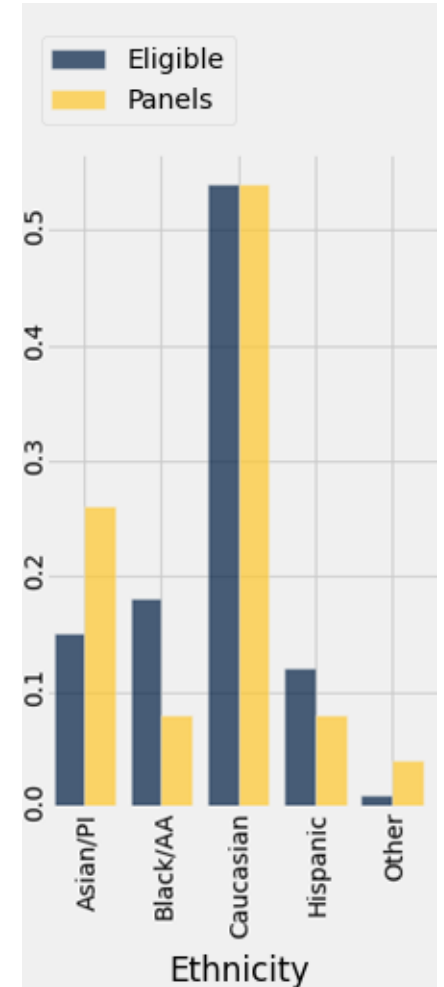
To run a hypothesis test we need to select a statistic

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 Alameda county is 0.28

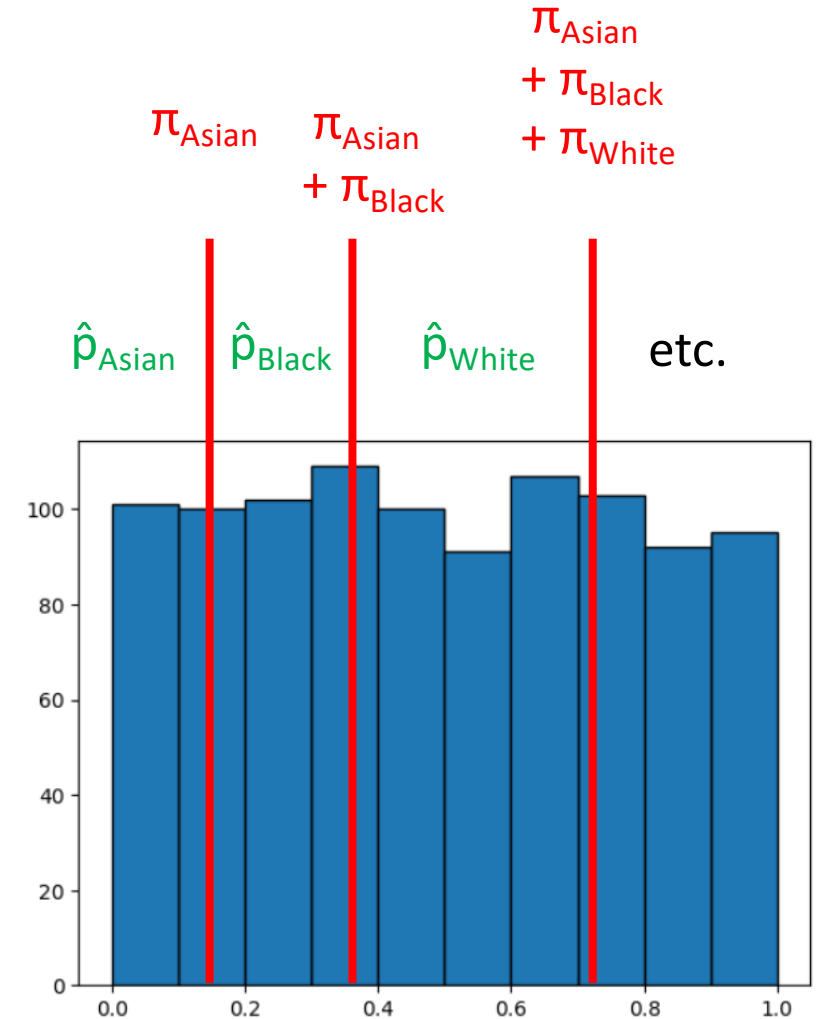


# Step 3: Creating a null distribution

To create a null distribution, we need to randomly generate several proportions consistent with the null hypothesis

- i.e.,  $\hat{p}_{\text{Asian}}$ ,  $\hat{p}_{\text{Black}}$  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

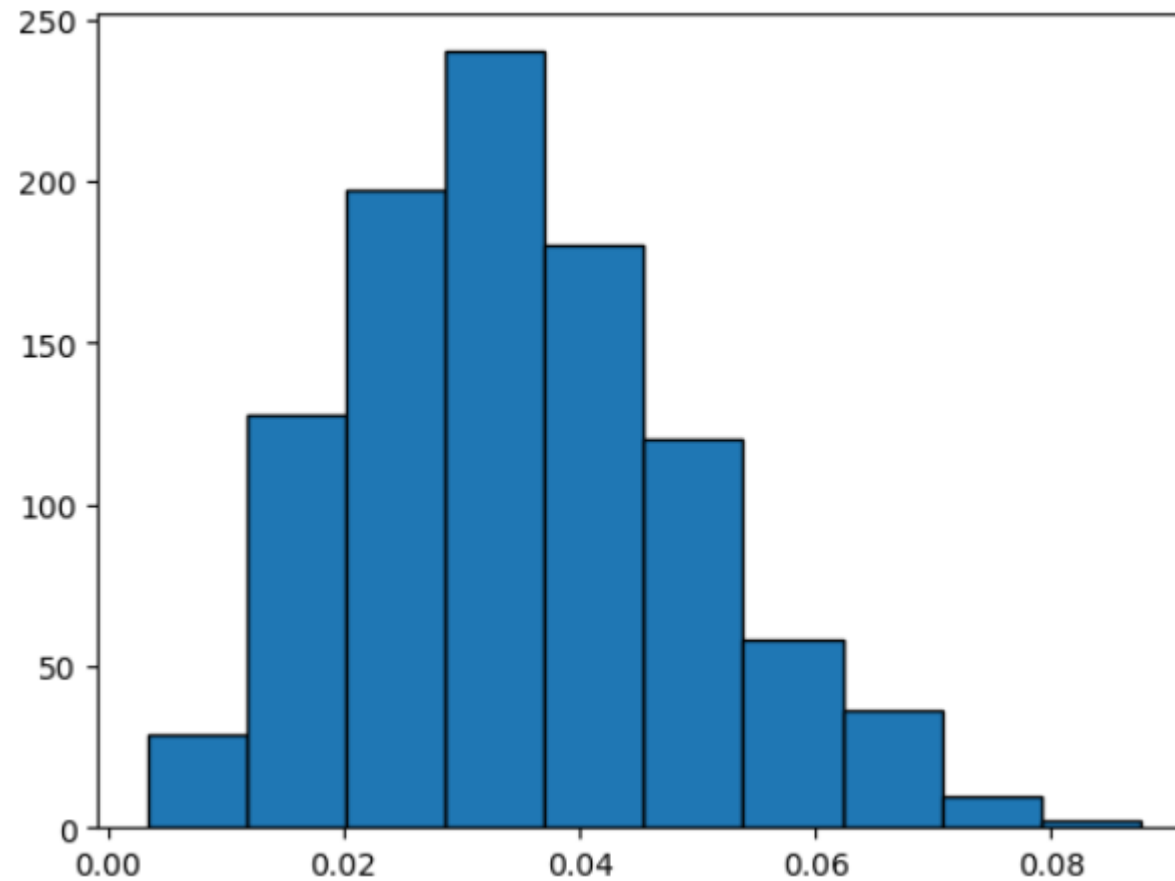
Once we have generated  $\hat{p}_{\text{Asian}}$ ,  $\hat{p}_{\text{Black}}$  etc. consistent with the null hypothesis, we can then calculate the TVD between these random and the true  $\hat{p}$ 's and the  $\pi_i$ 's specified by the null hypothesis

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

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



## Step 3: Creating a null distribution

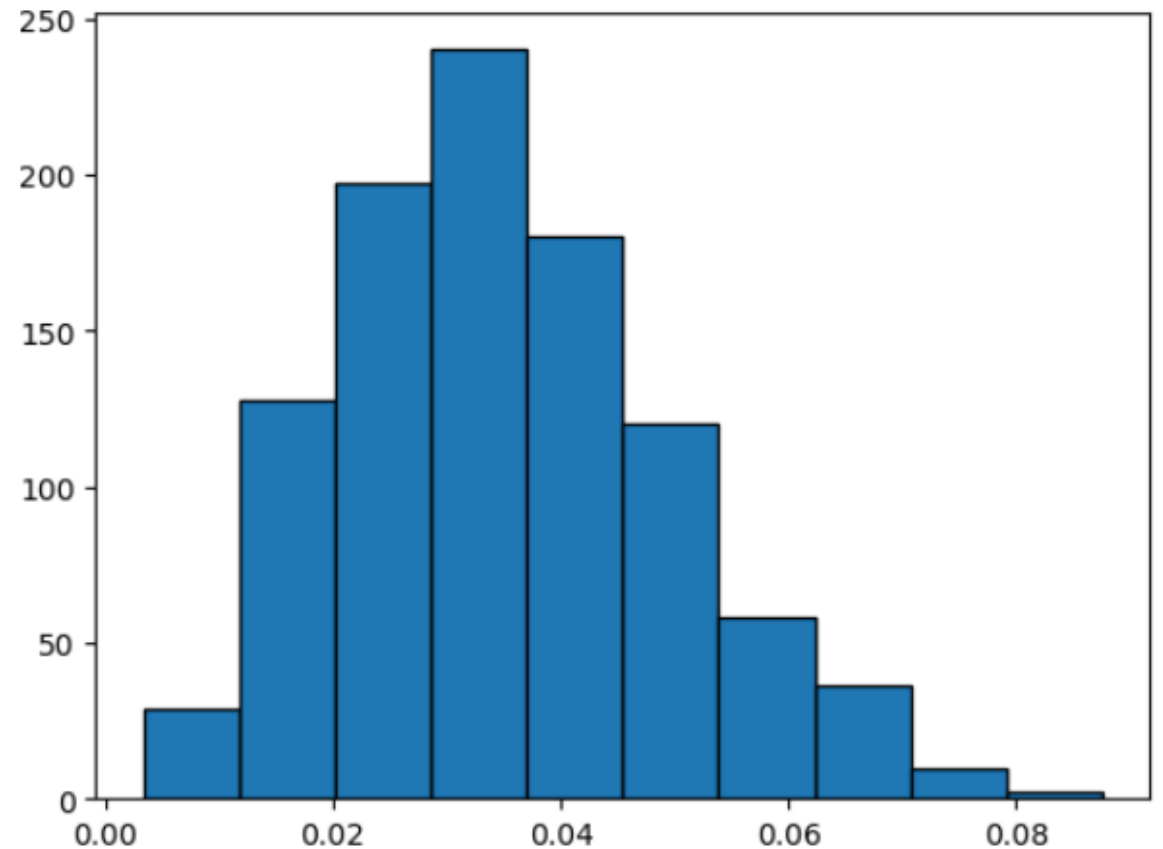


## Step 4: Calculate the p-value

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!

# Summary of the method

To assess whether a sample was drawn randomly from a known categorical distribution:

- Use TVD as the statistic because it measures the distance between categorical distributions
- Sample at random from the population and compute the TVD from the random sample; repeat numerous times
- Compare:
  - Empirical distribution of simulated TVDs
  - Actual TVD from the sample in the study