

YData: Introduction to Data Science



Class 17: Introduction to Statistical Inference

Overview

Review of for loops and functions

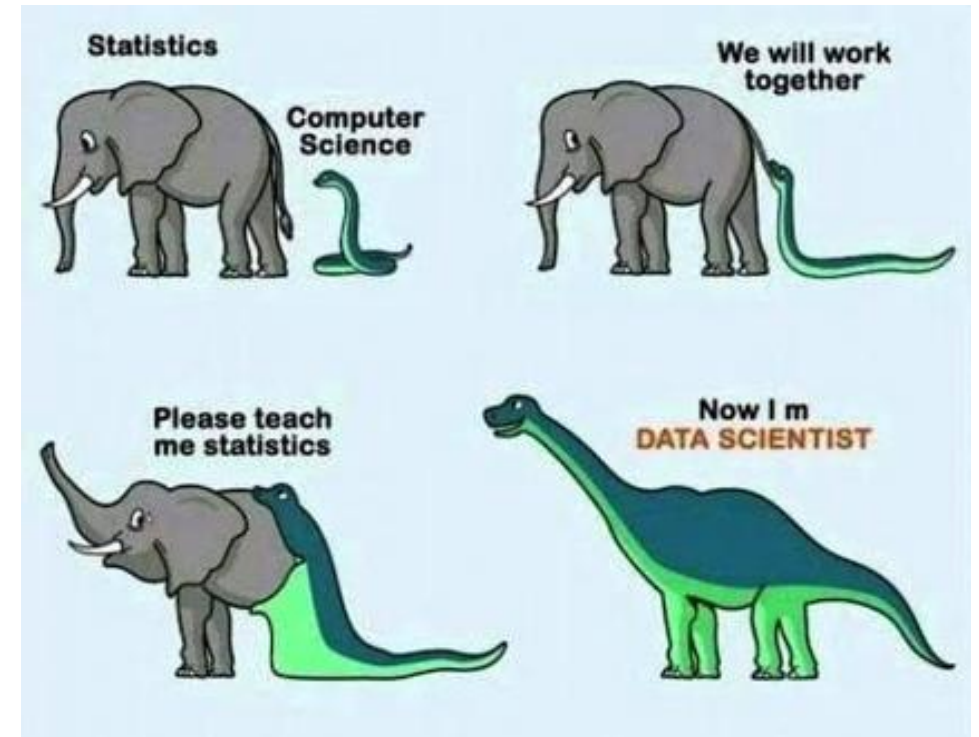
Central concepts in Statistical Inference

- Parameters and statistics

Sampling distributions

Hypothesis tests

- If there is time: Hypothesis tests for a single proportion



Reminder: keep working on your class project

A **polished** draft of the project is due on **November 10th**

Also, homework 7 is due on **Sunday November 3rd**

- I recommend finishing it early and then starting on your project by coming up with a topic and getting the relevant data.



Quick review of for loops and conditional statements

Review: for loops

For loops repeat a process many times, iterating over a sequence of items

- Often we are iterating over an array of sequential numbers

```
animals = ["cat", "dog", "bat"]
```

```
for creature in animals:
```

```
    print(creature)
```

```
for i in range(10):
```

```
    print(i**2)
```



Review: conditional statements

Conditional statements control the sequence of computations that are performed in a program

We use the keyword **if** to begin a conditional statement to only execute lines of code if a particular condition is met.

We can use **elif** to test additional conditions

We can use an **else** statement to run code if none of the if or elif conditions have been met.

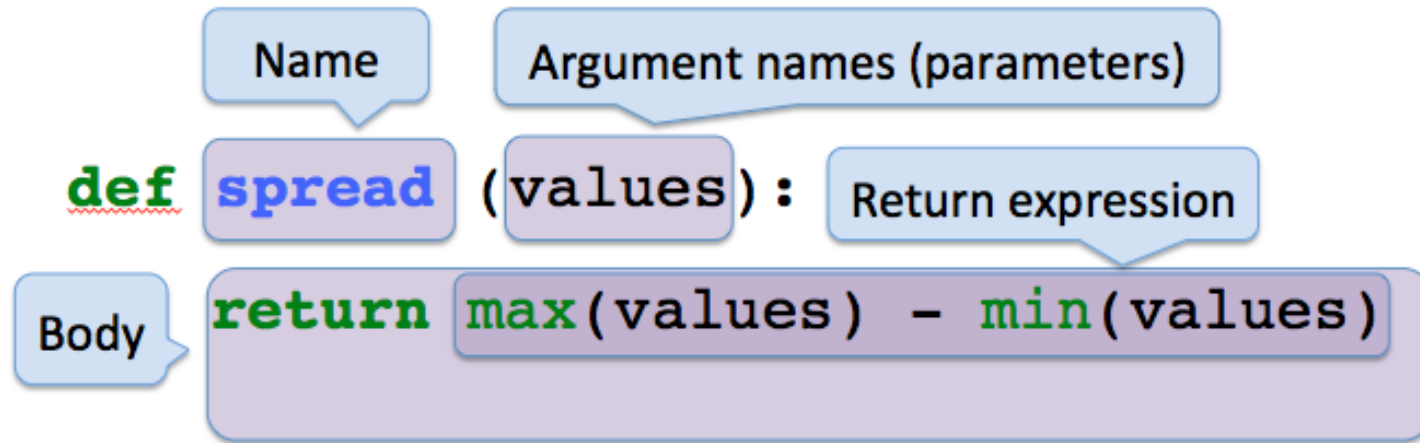
```
num = 5
if num == 1:
    print("Monday")
elif num == 2:
    print("Tuesday")
elif num == 3:
    print("Wednesday")
elif num == 4:
    print("Thursday")
elif num == 5:
    print("Friday")
elif num == 6:
    print("Saturday")
elif num == 7:
    print("Sunday")
else:
    print("Invalid input")
```

Let's do some warm up exercises in Jupyter!

Review of writing functions

Review of writing your own functions

User-defined functions give names to blocks of code



Functions can return tuples which allow us to return multiple names

```
val1, val2 = my_function()
```


Simulating flipping a coin

Let's practice writing functions by writing a function that can simulate flipping coins, where each coin has π probability of being heads

- Where π is a number between 0 and 1; e.g., $\pi = 0.5$ is a fair coin

We can do this using the following procedure:

1. Generate a random number between 0 and 1

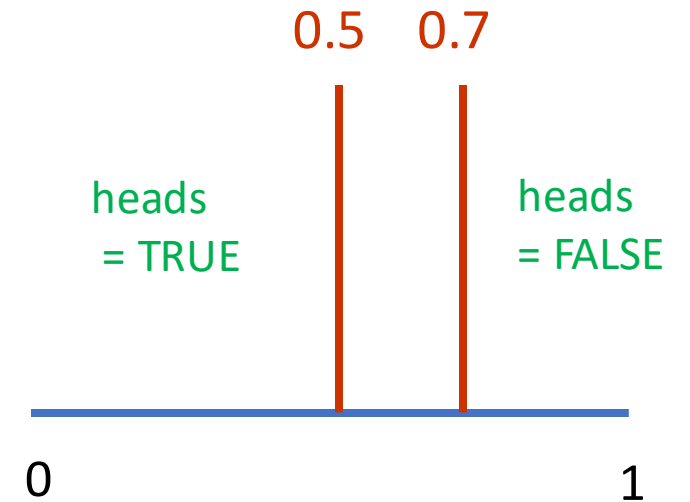
```
rand_num = np.random.rand(1)
```

2. Simulate a fair coin (.5) by mark values less than .5 as heads (True)

```
heads = rand_num <= .5
```

3. We can simulate a biased coin that will come up with heads 70% of the time ($\pi = 0.7$) using:

```
rand_num = np.random.rand(1)  
heads = rand_num <= .7
```



Simulating n random coin flips

We can simulate the number of heads we would get flipping a coin n times using:

1. Generate n random numbers uniformly distributed between 0 and 1

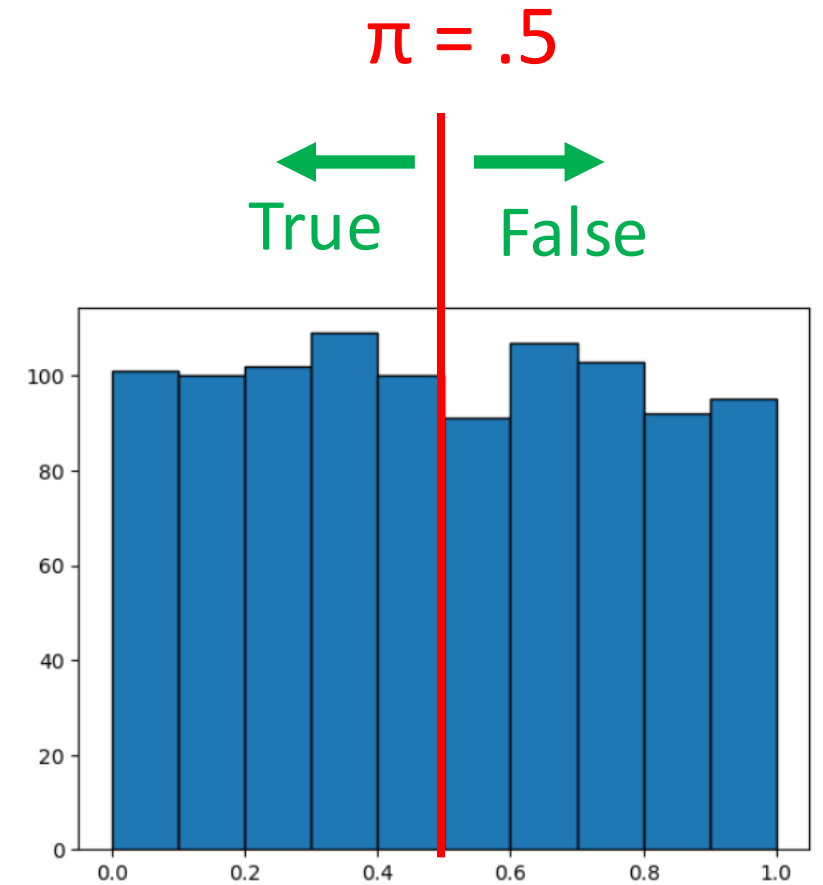
```
rand_nums = np.random.rand(n)
```

2. Mark points less than π as being **True**, and greater π than as being **False**

```
rand_binary = rand_nums <= prob_value
```

3. Sum the number of heads (**True's**) we get

```
num_heads = np.sum(rand_binary)
```



Let's explore this in Jupyter!

Statistical Inference

Statistical Inference

In **statistical inference** we use a sample of data to make claims about a larger population (or process)

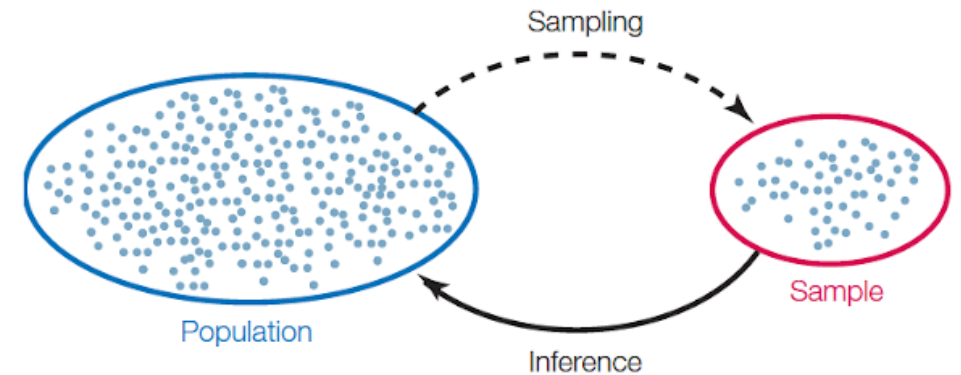
Examples:

Voting data

- **Population:** everyone who will vote
- **Sample:** survey of 1,000 randomly selected voters

Bechdel data:

- **Population:** All movies with budgets > \$10,000,000
- **Sample:** 1794 movies randomly selected



Population: all individuals/objects of interest

Sample: A subset of the population

Terminology

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

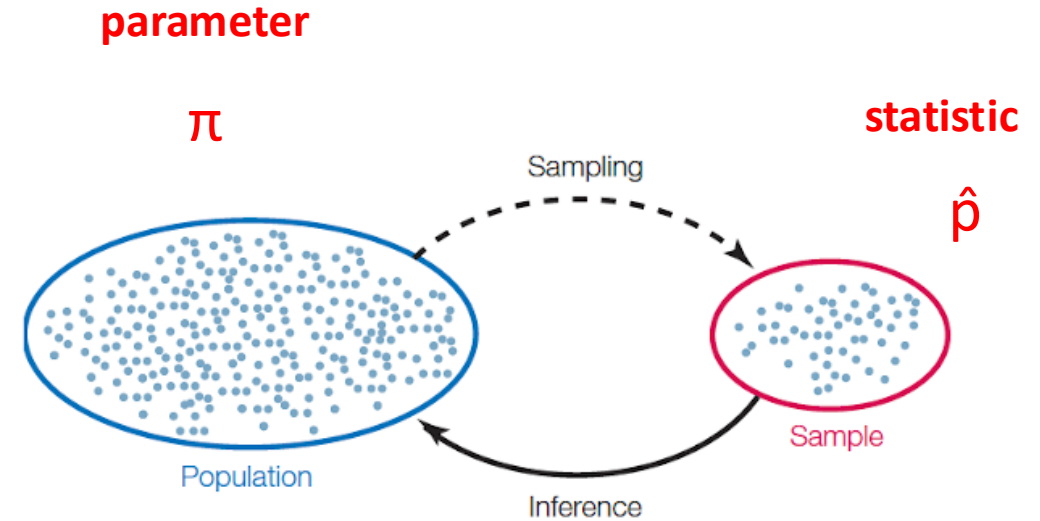
- e.g., population proportion π
- e.g., the proportion of voters who will vote for Trump

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

- e.g., sample proportion \hat{p}
- e.g., proportion of 1,000 people in our sample

A statistic can be used as an estimate of a parameter

- A parameter is a single fixed value
- Statistics tend to vary from sample to sample



Example:

- Using the proportion of 1,000 voters (\hat{p}_{Trump}) to estimate the proportion of all voters who will vote for Trump (π_{Trump})

Examples of parameters and statistics

	Sample statistic	Population parameter	Bechdel example
Mean	\bar{x}	μ	statistics. <code>mean</code> (domgross_2013) $\bar{x} = \$95,174,783$
Proportion	\hat{p}	π	bechdel. <code>count</code> ("PASS")/len(bechdel) $\hat{p} = 0.45$
Correlation	r	ρ	statistics. <code>correlation</code> (budget_2013, domgross_2013) $r = 0.46$

Sampling

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

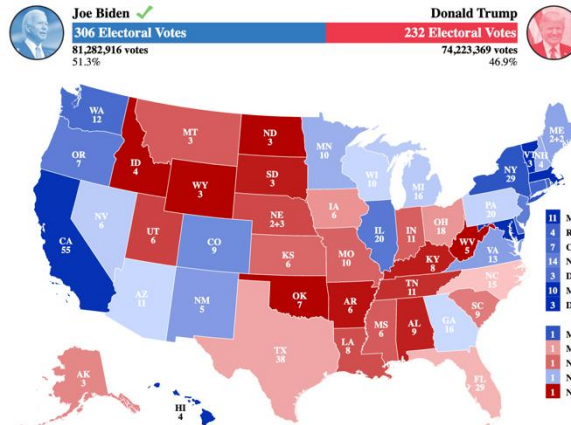
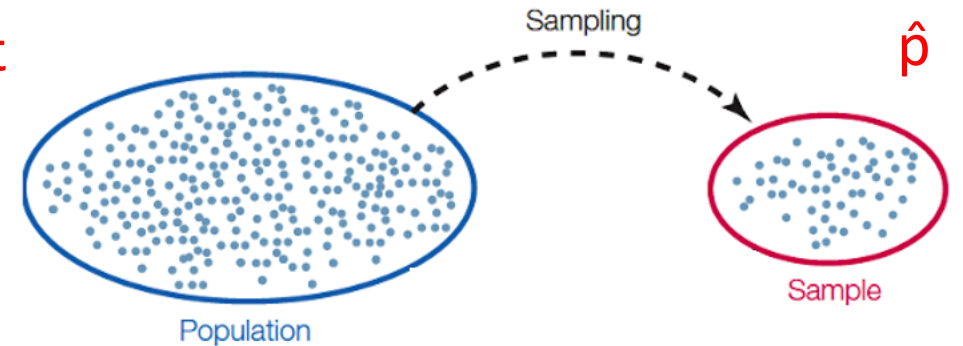
- Allows for generalizations to the population

parameter

π

statistic

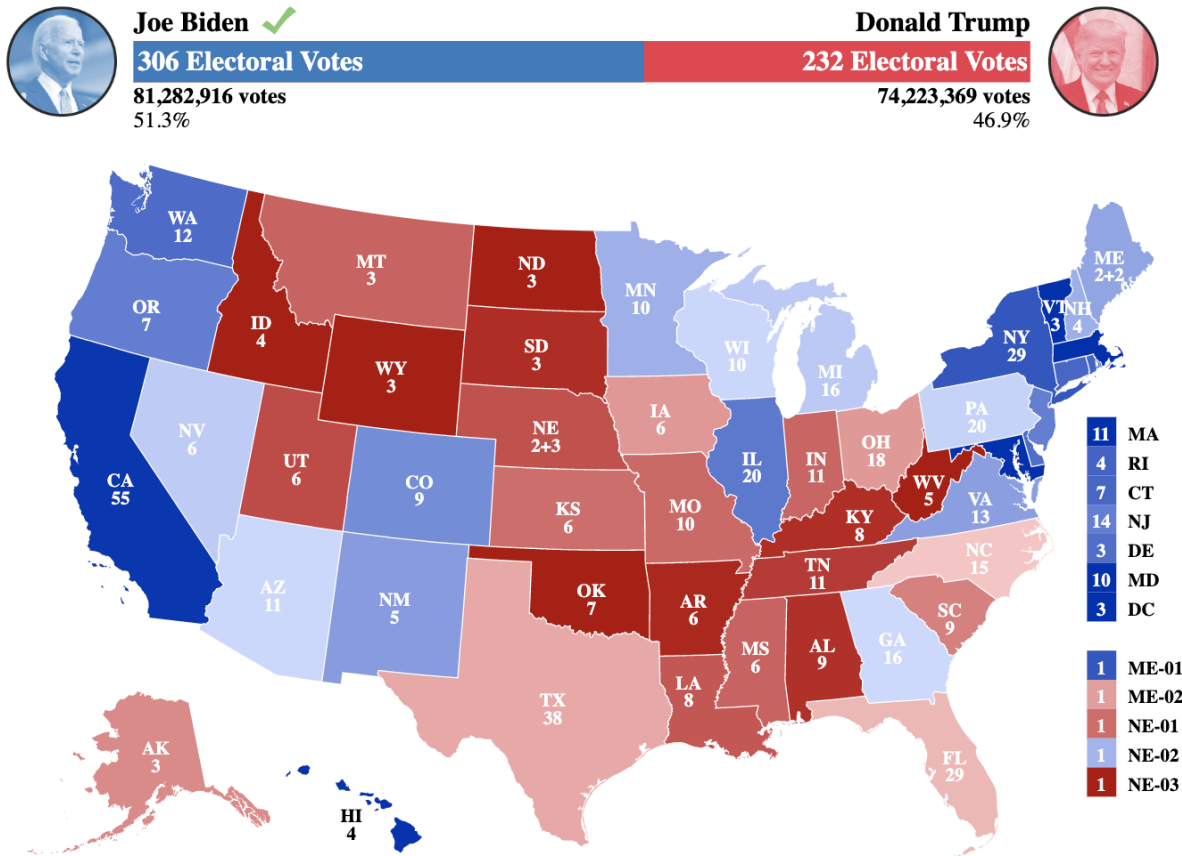
\hat{p}



Polls of 1,000 voters: \hat{p}_{Trump}

Vote on election day: π_{Trump}

Example: The 2020 US Presidential Election



According to The Cook Political Report, the voting outcome in Georgia in 2020 was

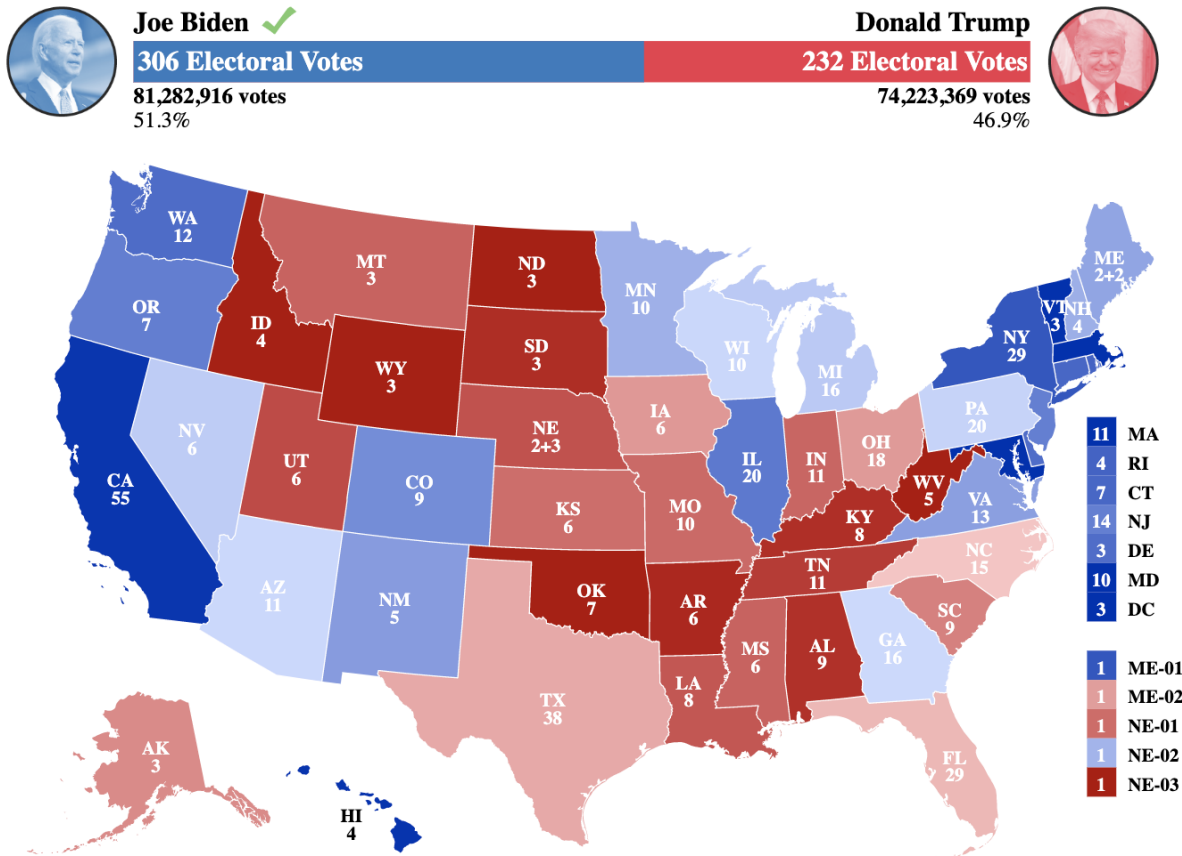
- Trump = 2,461,854
- Biden = 2,473,633

We can denote the proportion of the vote that Trump got using

π_{Trump}

- Q: what is the value of π_{Biden} ?

Example: The 2020 US Presidential Election



If 1,000 voters were randomly sampled, we could denote the proportion in the sample that voted for Biden using: \hat{p}_{Biden}

Would we expect \hat{p}_{Biden} to be equal to π_{Biden} ?

If we repeated the process of sampling another 1,000 random voters, would we expect to get the same \hat{p}_{Biden} ?

Let's explore this in Jupyter!

Sampling distributions

Probability distribution of a statistic

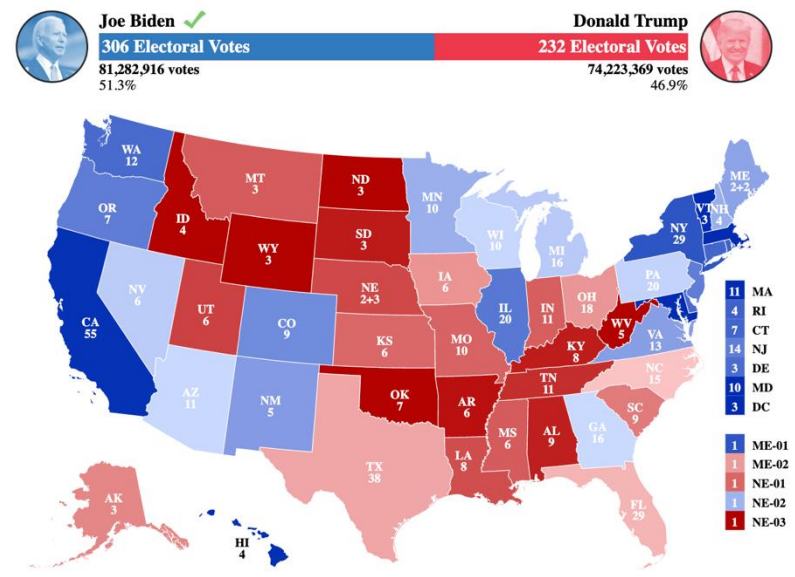
Values of a statistic vary because random samples vary

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 a simulated statistics

π_{Trump}

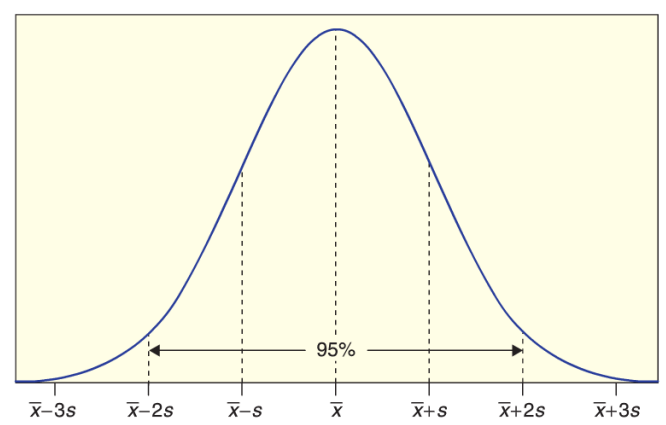
n = 1,000



\hat{p}_{Trump}



\hat{p}_{Trump}



Sampling distribution!



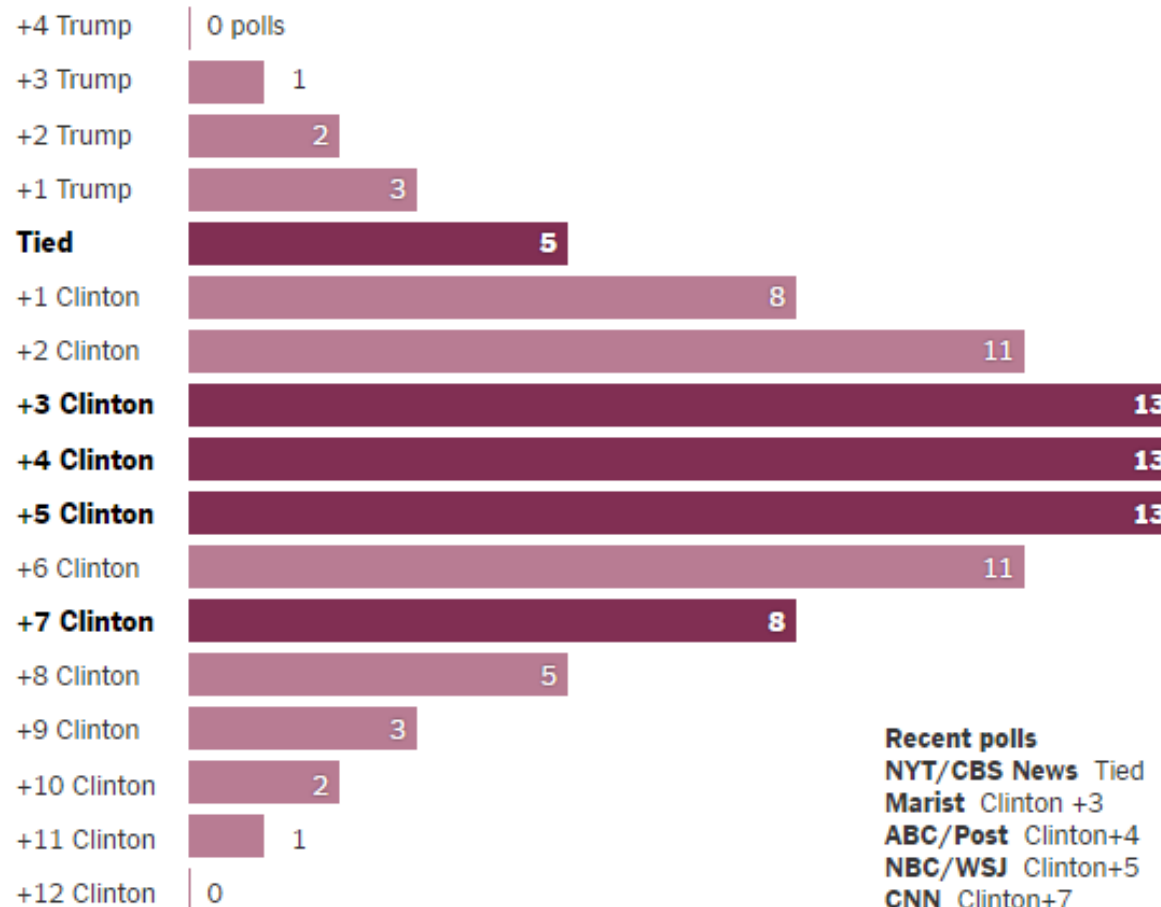
\hat{p}_{Trump}

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?

Let's explore this in Jupyter!

Simulating flipping a coin

We can simulate flipping a fair coin using the following procedure

1. Generated a random number between 0 and 1

```
rand_num = np.random.rand(1)
```

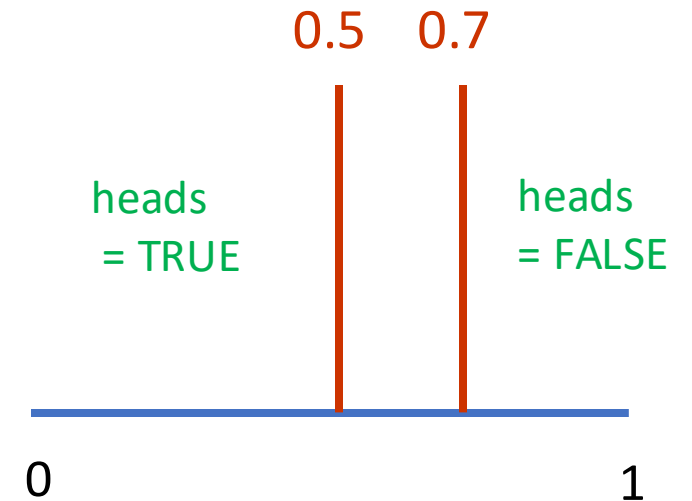
2. We mark values less than .5 has heads (True)

```
heads = rand_num <= .5
```

We can simulate a biased coin that will come up with heads 70% of the time using

```
rand_num = np.random.rand(1)
```

```
heads = rand_num <= .7
```



Simulating a random proportion (\hat{p})

We can simulate a random proportions \hat{p} (from a sample of size n) 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 \hat{p}

```
rand_phat = np.mean(rand_binary)
```



Let's explore this in Jupyter!

Hypothesis tests

A quick note on probability

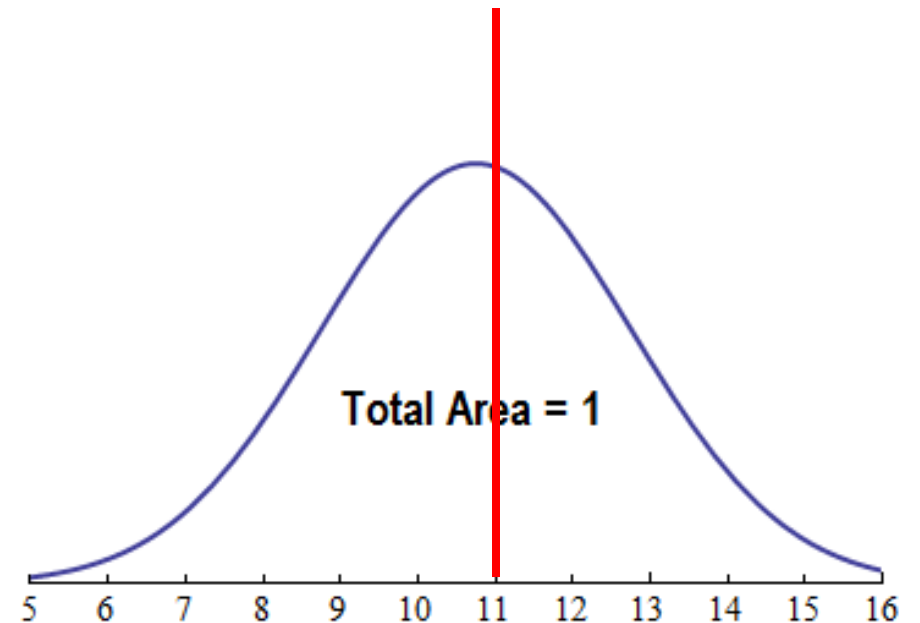
Probability is a way of measuring the likelihood that an event will occur

Probability models assigns a number between 0 and 1 to the outcome of an event (outcome) occurring

We can use a probability model to calculate the probability of an event

For example:

- $P(X < 11) = 0.55$
- $P(X > 20) = 0$



Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population (parameter)

Example 1: The average body temperature of humans is 98.6°

How can we write this using symbols?

- $\mu = 98.6$

Statistical tests (hypothesis test)

A **statistical test** uses data from a sample to assess a claim about a population (parameter)


Example 2: A higher proportion of voters will vote for Trump compared to Harris

How can we write this using symbols?

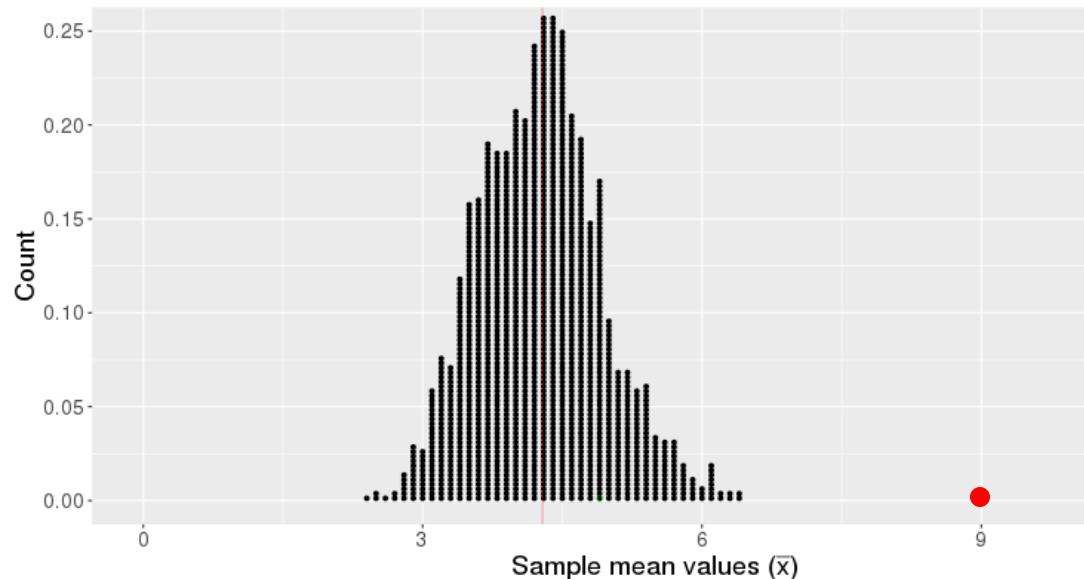
- $\pi_{\text{Trump}} > \pi_{\text{Harris}}$ or $\pi_{\text{Trump}} - \pi_{\text{Harris}} > 0$

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



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?

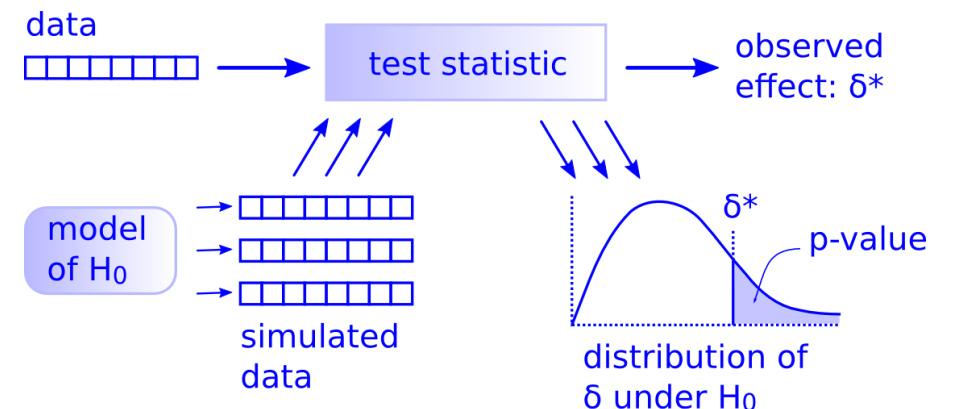
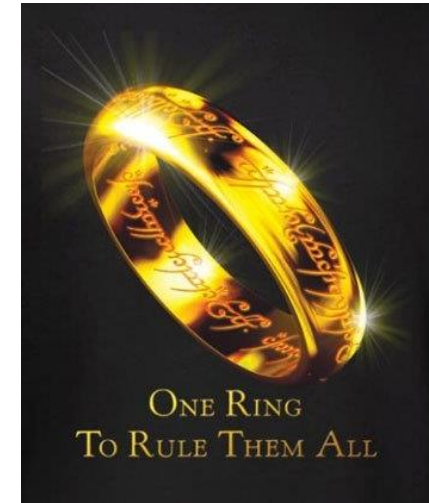
	title	binary
1	Dredd 3D	PASS
2	12 Years a Slave	FAIL
3	2 Guns	FAIL
4	42	FAIL
5	47 Ronin	FAIL
6	A Good Day to Die Hard	FAIL
7	About Time	PASS
8	Admission	PASS
9	After Earth	FAIL
10	American Hustle	PASS
11	August: Osage County	PASS
12	Beautiful Creatures	PASS
13	Blue Jasmine	PASS
14	Captain Phillips	FAIL

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 only 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?

$$H_A: \pi < 0.5$$

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

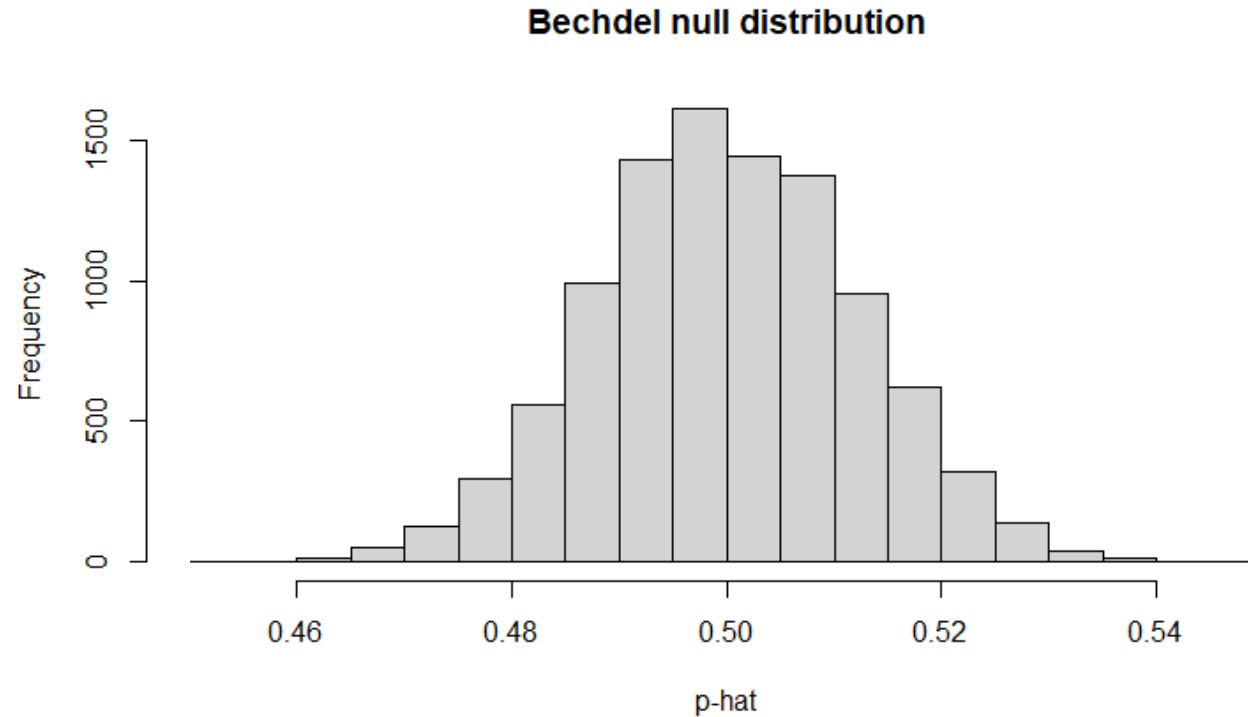
Let's simulate flipping a coin 1794 times and see how many times we get 803 **or fewer** heads

Chance models

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

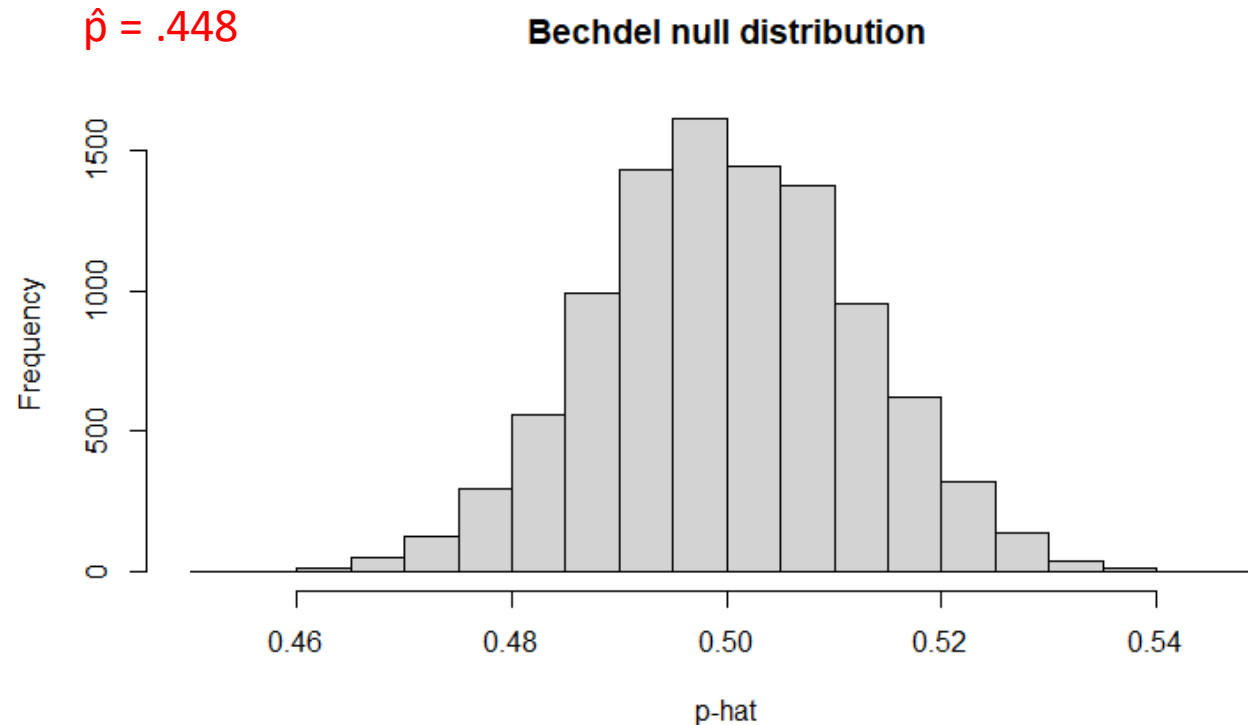
Any ideas how to do this?

Simulating Flipping 1794 coins 10,000 times



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

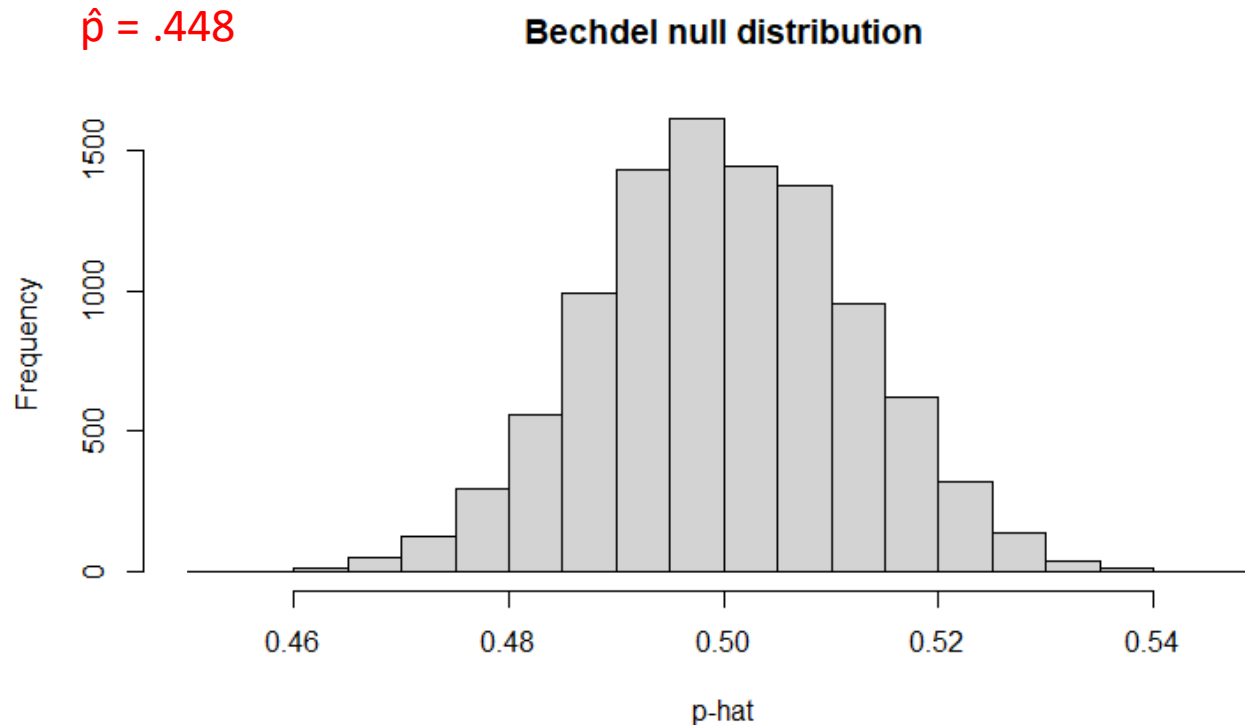
Step 4: calculate the p-value



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

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

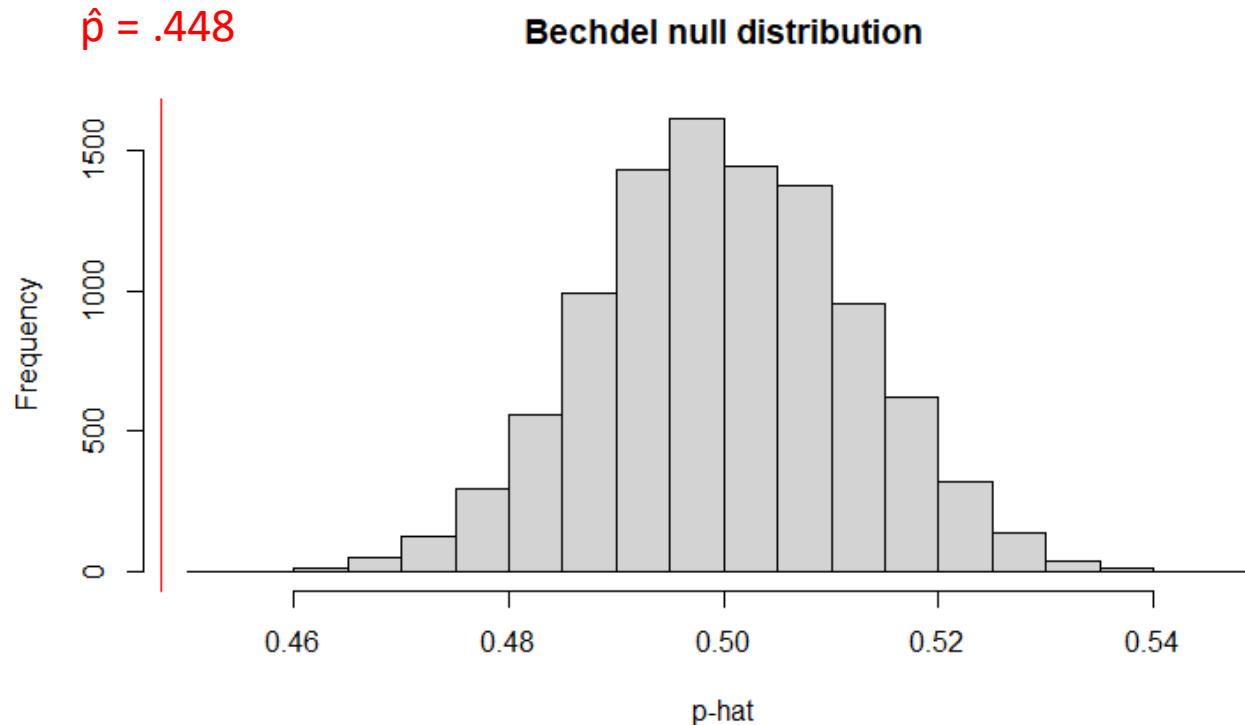
Step 4: calculate the p-value



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

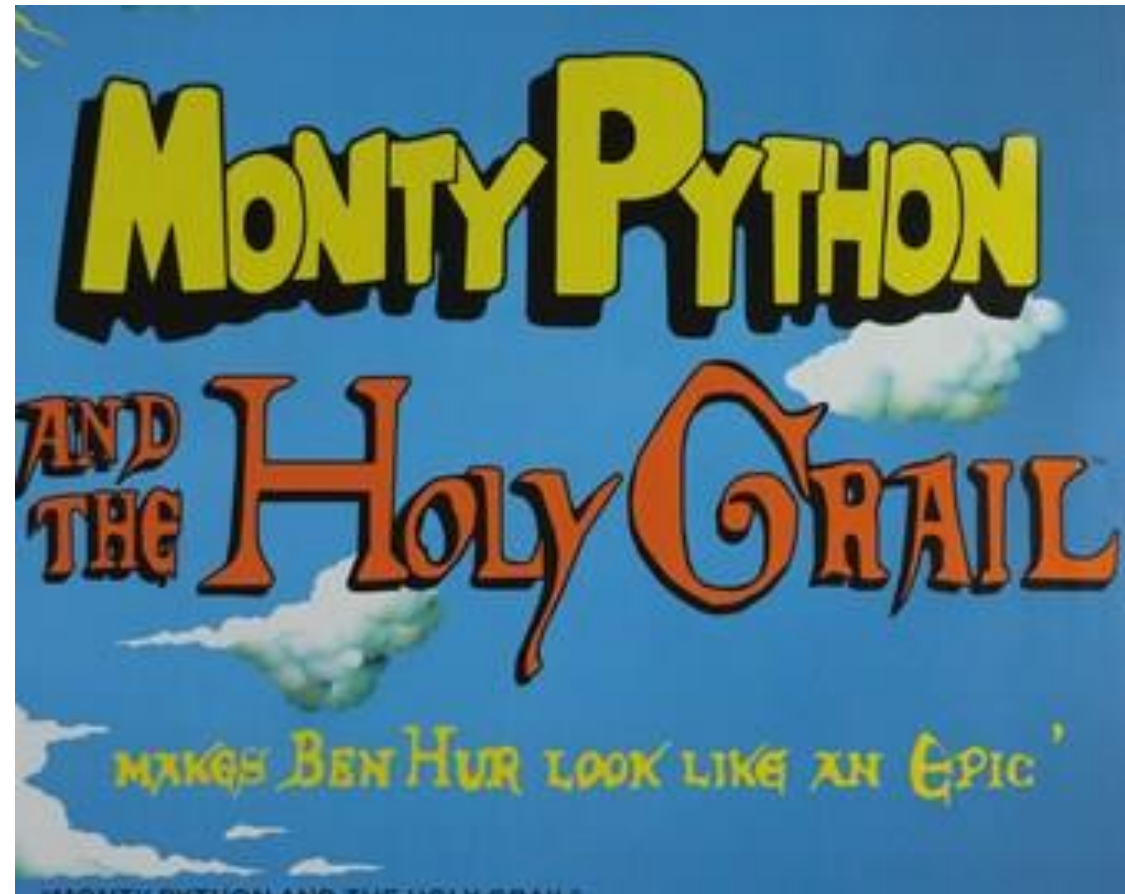
Step 5: Make a decision



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



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

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'



$$\hat{p} = .448$$

