YData: Introduction to Data Science

Class 19: Introduction to Statistical Inference

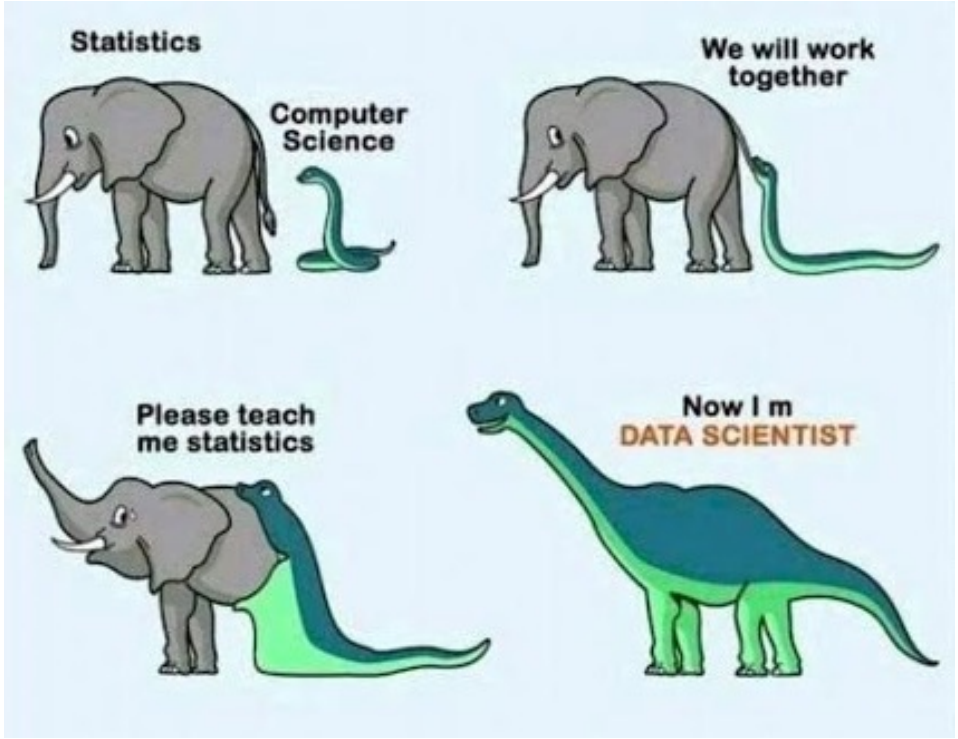
Overview

Very quick review/continuation of maps

Introduction to Statistical Inference

- Parameters and statistics

Introduction to hypothesis tests



Reminder: class project

The final project is a 6-10 page Jupyter notebook report where you analyze your own data to address a question that you find interesting

- A project template Jupyter notebooks is on Canvas

A polished draft of the project is due on April 11th

- 2 day extension because of Easter

Focus on giving insight into some interesting questions

- You do not need to use all methods discussed in the class

Project timeline

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 will be posted to Canvas

Wednesday, April 10th

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

Sunday, April 30th

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

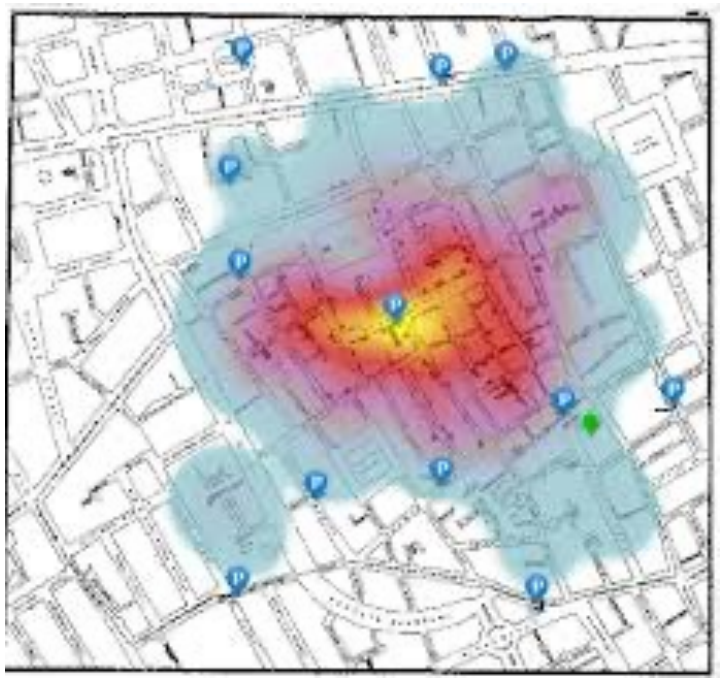
Quick review of mapping

Quick review of maps

Visualizing data on a map can be a powerful way to see spatial trends

We can create maps in Python using geopandas DataFrames

- Like regular DataFrames with an additional geometry column that has Shaply objects



John Snow's ghost map (1854)

	key_comb_drvr	geometry
0	M11551	POINT (117.525391 34.008926)
1	M17307	POINT (86.51248 30.474344)
2	M19584	POINT (89.537415 37.157627)

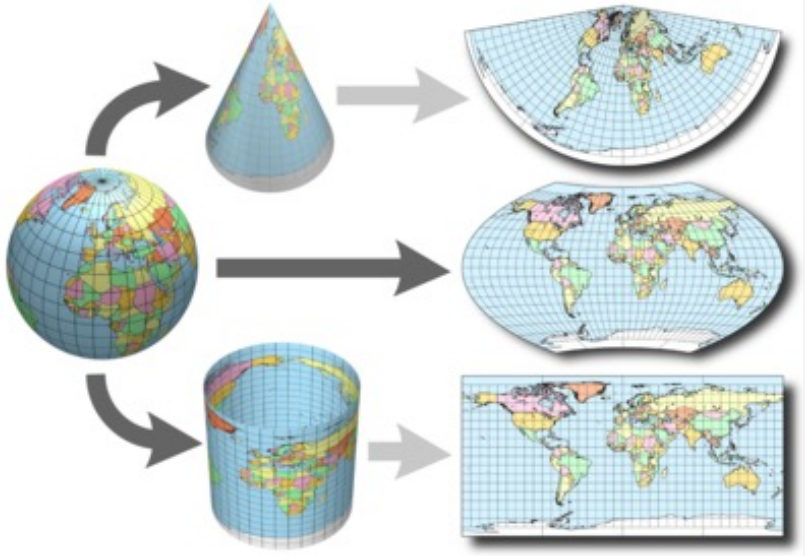
Review: CRSs and map projections

A coordinate reference system (CRS) is a framework used to precisely measure locations on the surface of the Earth as coordinates

- Needed for aligning different layers on maps

There are many map projections to display Earth's 3D structure on a 2D map surface.

- **Mercator projection** keeps angles intact
- **Eckert IV projection** keeps the size of land areas intact



WHAT YOUR FAVORITE MAP PROJECTION SAYS ABOUT YOU

MERCATOR

YOU'RE NOT REALLY INTO MAPS.

VAN DER GRINTEN

YOU'RE NOT A COMPLICATED PERSON. YOU LOVE THE MERCATOR PROJECTION; YOU JUST WISH IT WEREN'T SQUARE. THE EARTH'S NOT A SQUARE, IT'S A CIRCLE. YOU LIKE CIRCLES. TODAY IS GONNA BE A GOOD DAY!

HOBO-DYER

YOU WANT TO AVOID CULTURAL IMPERIALISM, BUT YOU'VE HEARD BAD THINGS ABOUT GALL-PETERS. YOU'RE CONFLICT-AVERSE AND BUY ORGANIC. YOU USE A RECENTLY-INVENTED SET OF GENDER-NEUTRAL PRONOUNS AND THINK THAT WHAT THE WORLD NEEDS IS A REVOLUTION IN CONSCIOUSNESS.

PLATE CARRÉE (EQUIRECTANGULAR)

YOU THINK THIS ONE IS FINE. YOU LIKE HOW X AND Y MAP TO LATITUDE AND LONGITUDE. THE OTHER PROJECTIONS OVERCOMPLICATE THINGS. YOU WANT ME TO STOP ASKING ABOUT MAPS SO YOU CAN ENJOY DINNER.

ROBINSON

YOU HAVE A COMFORTABLE PAIR OF RUNNING SHOES THAT YOU WEAR EVERYWHERE. YOU LIKE COFFEE AND ENJOY THE BEATLES. YOU THINK THE ROBINSON IS THE BEST-LOOKING PROJECTION, HANDS DOWN.

DYMAXION

YOU LIKE ISAAC ASIMOV, XML, AND SHOES WITH TOES. YOU THINK THE SEGWAY GOT A BAD RAP. YOU OWN 3D GOGGLES, WHICH YOU USE TO VIEW ROTATING MODELS OF BETTER 3D GOGGLES. YOU TYPE IN DVORAK.

A GLOBE!

YES, YOU'RE VERY CLEVER.

WATERMAN BUTTERFLY

REALLY? YOU KNOW THE WATERMAN? HAVE YOU SEEN THE 1909 CAHILL MAP IT'S BASED— ... YOU HAVE A FRAMED REPRODUCTION AT HOME?! WHOA. ... LISTEN, FORGET THESE QUESTIONS. ARE YOU DOING ANYTHING TONIGHT?

WINKEL-TRIPEL

NATIONAL GEOGRAPHIC ADOPTED THE WINKEL-TRIPEL IN 1998, BUT YOU'VE BEEN A WT FAN SINCE *LONG* BEFORE 'NATGEO' SHOWED UP. YOU'RE WORRIED IT'S GETTING PLAYED OUT, AND ARE THINKING OF SWITCHING TO THE KAVRAYSKIY. YOU ONCE LEFT A PARTY IN DISGUST WHEN A GUEST SHOWED UP WEARING SHOES WITH TOES. YOUR FAVORITE MUSICAL GENRE IS "POST-".

GOODE HOMOLOSINE

THEY SAY MAPPING THE EARTH ON A 2D SURFACE IS LIKE FLATTENING AN ORANGE PEEL, WHICH SEEMS EASY ENOUGH TO YOU. YOU LIKE EASY SOLUTIONS. YOU THINK WE WOULDN'T HAVE SO MANY PROBLEMS IF WE'D JUST ELECT *NORMAL* PEOPLE TO CONGRESS INSTEAD OF POLITICIANS. YOU THINK AIRLINES SHOULD JUST BUY FOOD FROM THE RESTAURANTS NEAR THE GATES AND SERVE *THAT* ON BOARD. YOU CHANGE YOUR CAR'S OIL, BUT SECRETLY WONDER IF YOU REALLY *NEED* TO.

PEIRCE QUINCUNCIAL

YOU THINK THAT WHEN WE LOOK AT A MAP, WHAT WE REALLY SEE IS OURSELVES. AFTER YOU FIRST SAW *INCEPTION*, YOU SAT SILENT IN THE THEATER FOR SIX HOURS. IT FREAKS YOU OUT TO REALIZE THAT EVERYONE AROUND YOU HAS A SKELETON INSIDE THEM. YOU *HAVE* REALLY LOOKED AT YOUR HANDS.

GALL-PETERS

I *HATE* YOU.

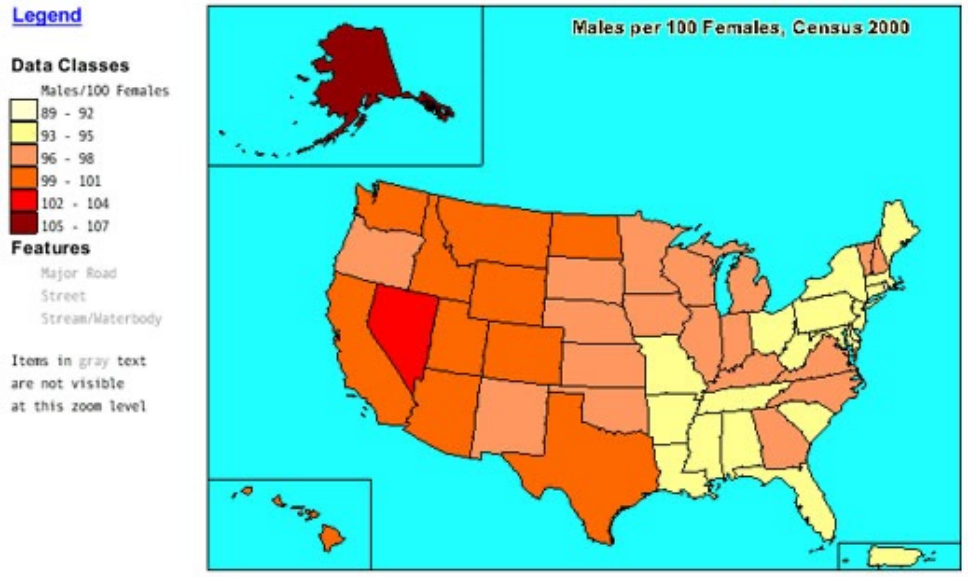
Review: Choropleth and Isopleth maps

Choropleth maps: shades/colors in predefined areas based on properties of a variable

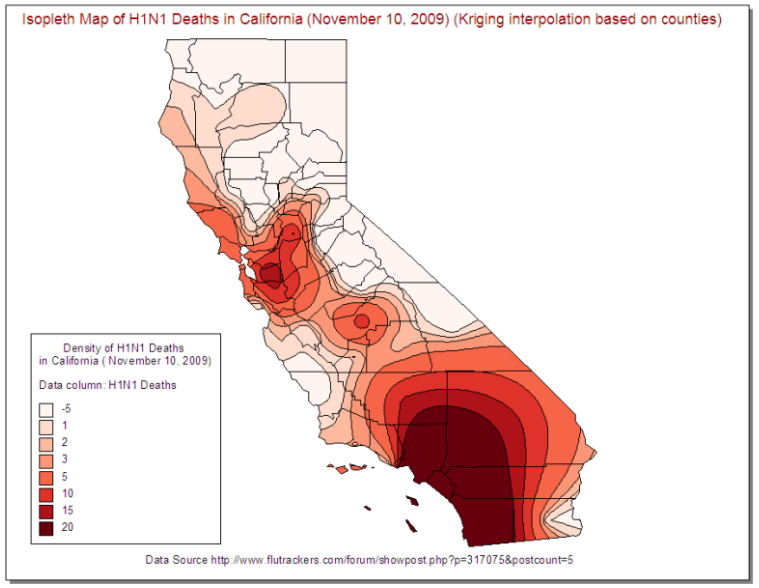
- We can then use the gpd.plot(column =) method to create choropleth maps

Isopleth maps: creates regions based on constant values

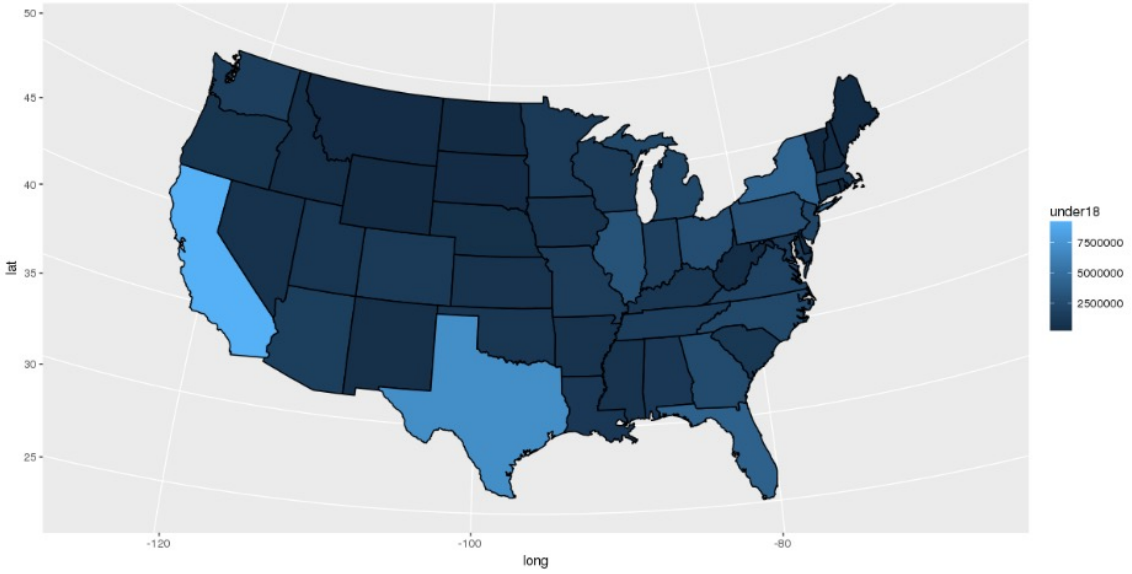
Choropleth map



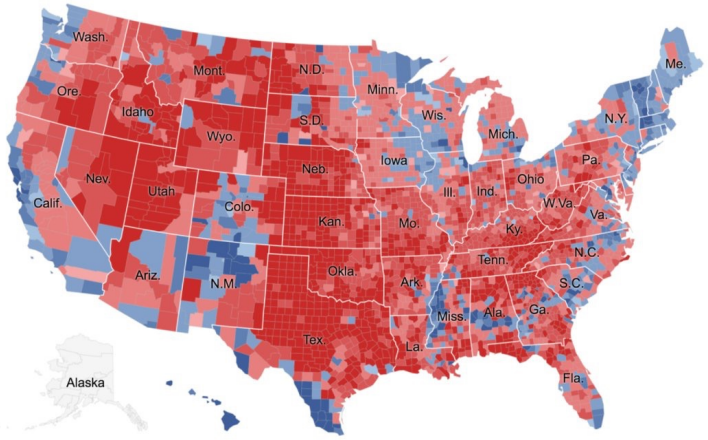
Isopleth map

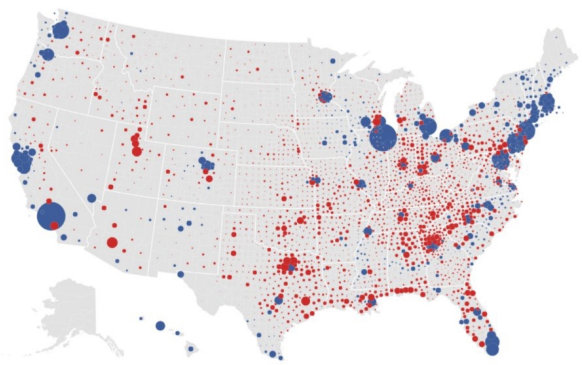


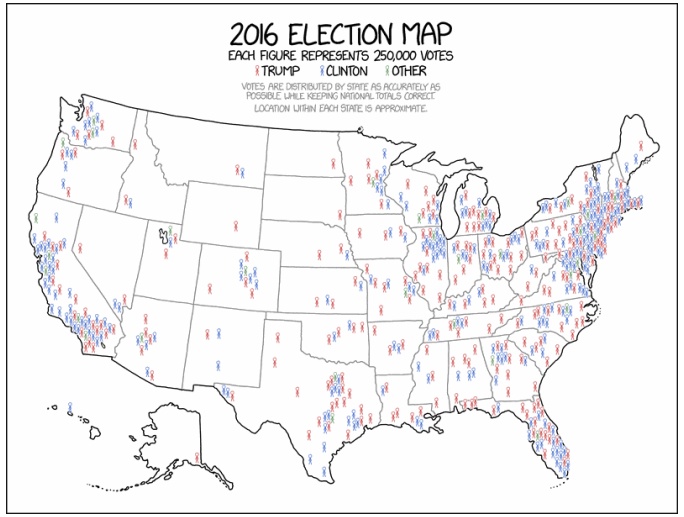
Review: Pet Peeve #208



Cloropleth maps can be misleading







Looks like most of the country voted republican

Let's look at a brief demo in Jupyter!

Statistical Inference

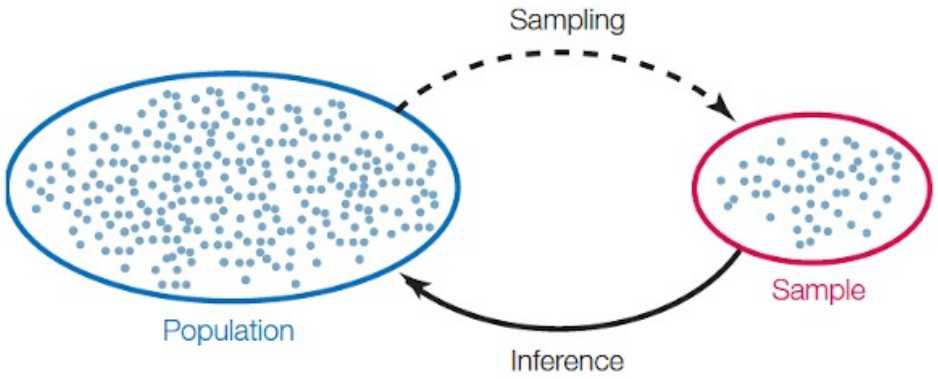
Inference

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

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

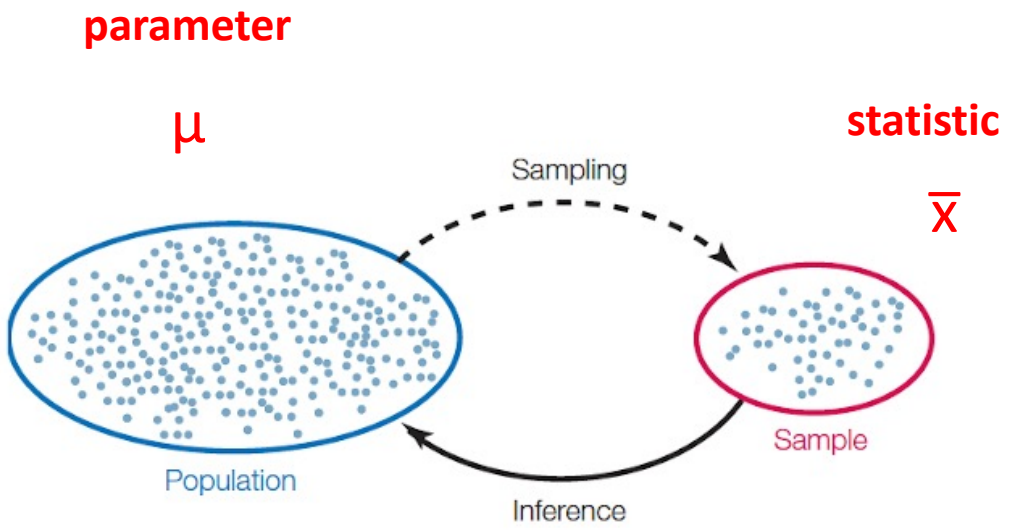
A parameter is number associated with the population

- e.g., population mean μ
- e.g., average height of all humans

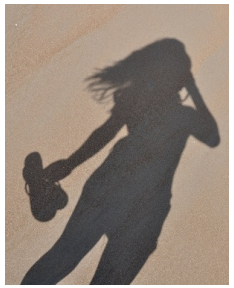
A **statistic** is number calculated from the sample

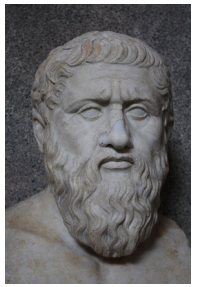
- e.g., sample mean $\bar{x}$
- e.g., average height of 1,000 people in our sample

A statistic can be used as an estimate of a parameter



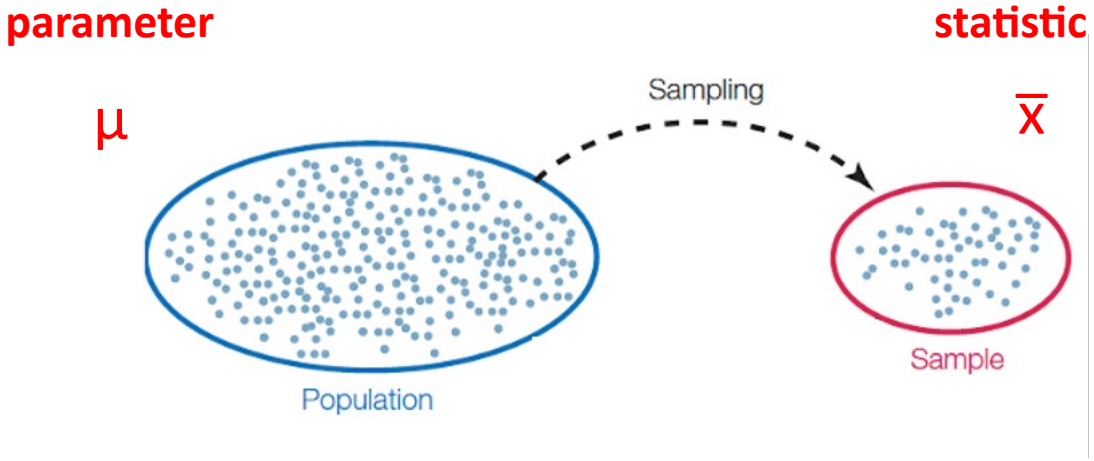
Examples of parameters and statistics





	Sample Statistic	**Population Parameter**
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Proportion	$\hat{p}$	π
Correlation	r	ρ
regression slope	b	β

Sampling



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

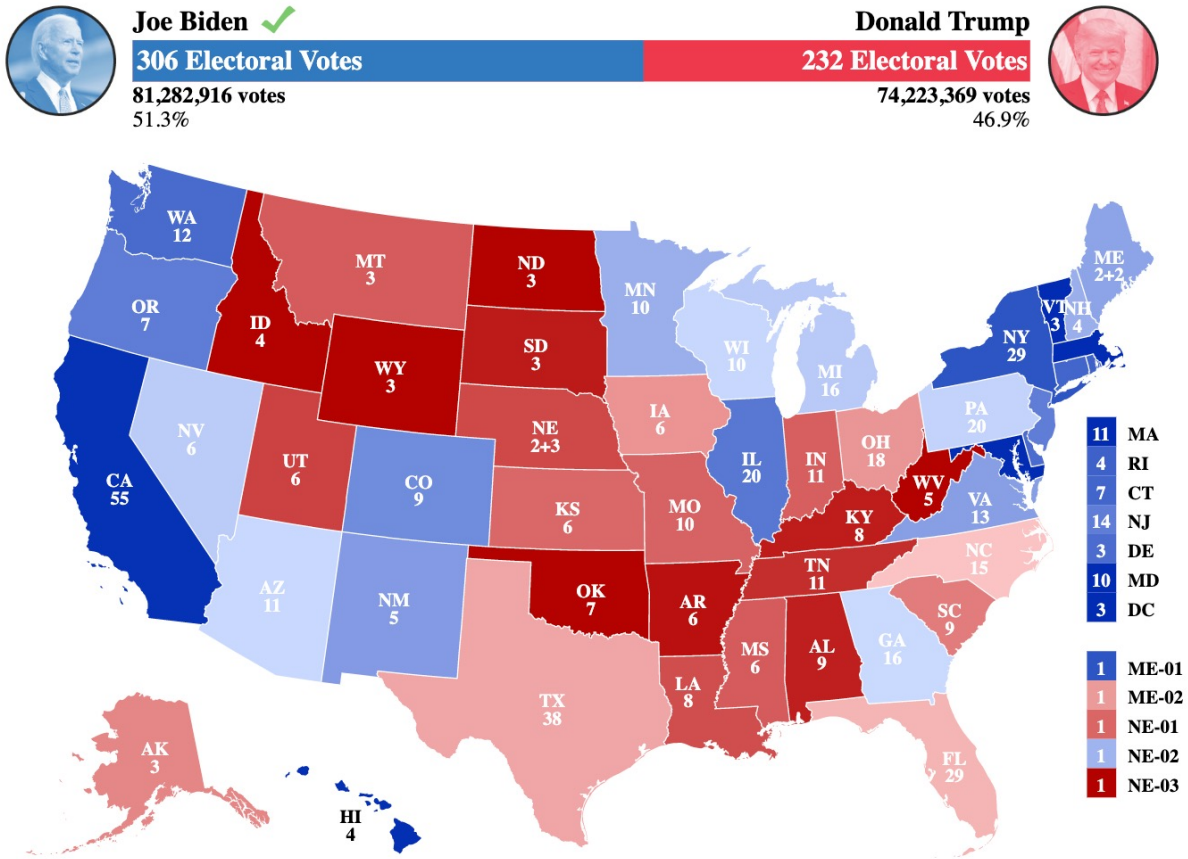
Allows for generalizations to the population!

- **No bias**: statistics (on average) equal parameter value

Why does this work?

- Soup analogy!

Example: The 2020 US Presidential Election



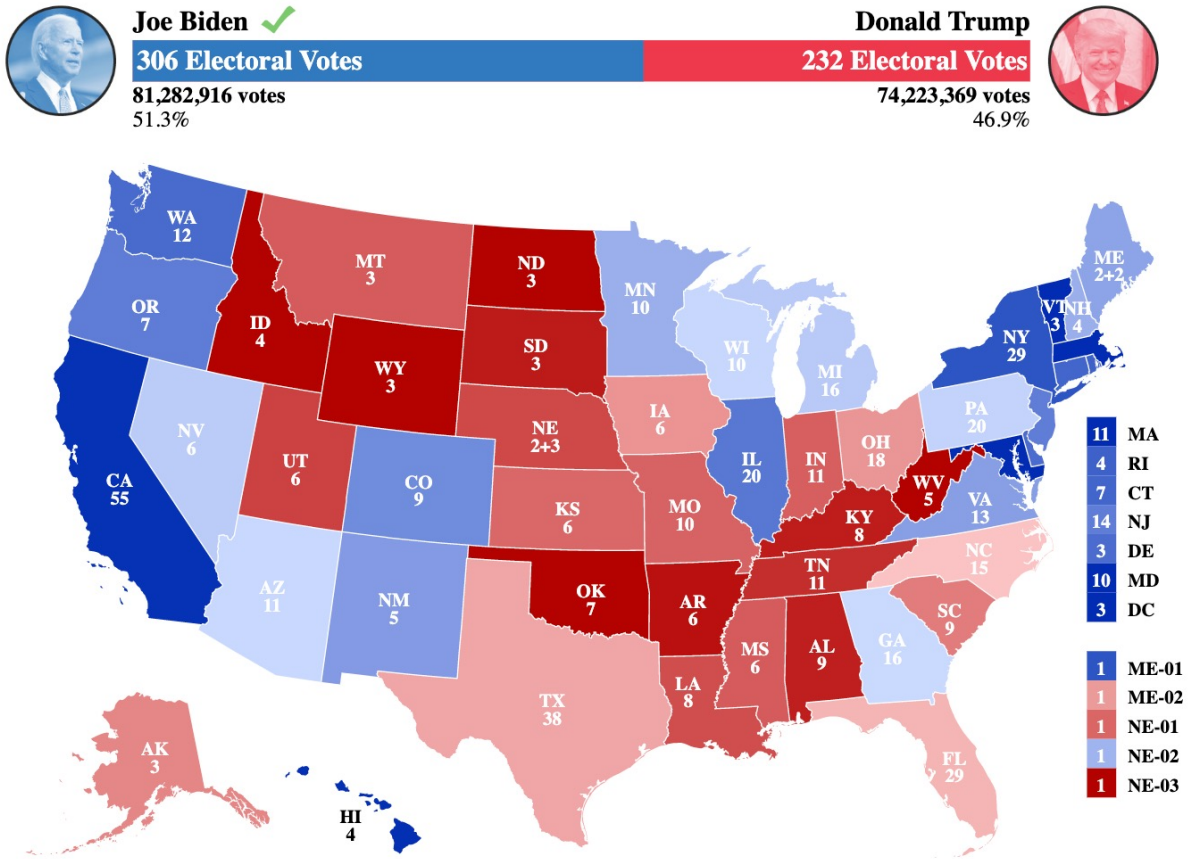
According to The Cook Political Report, the voting outcome in Georgia was

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

We can denote the proportion of the vote that Biden got using π_{Biden}

- 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 $\hat{}_{Biden}$?

Let's explore this in Jupyter!

Sampling distributions

Probability distribution of a statistic

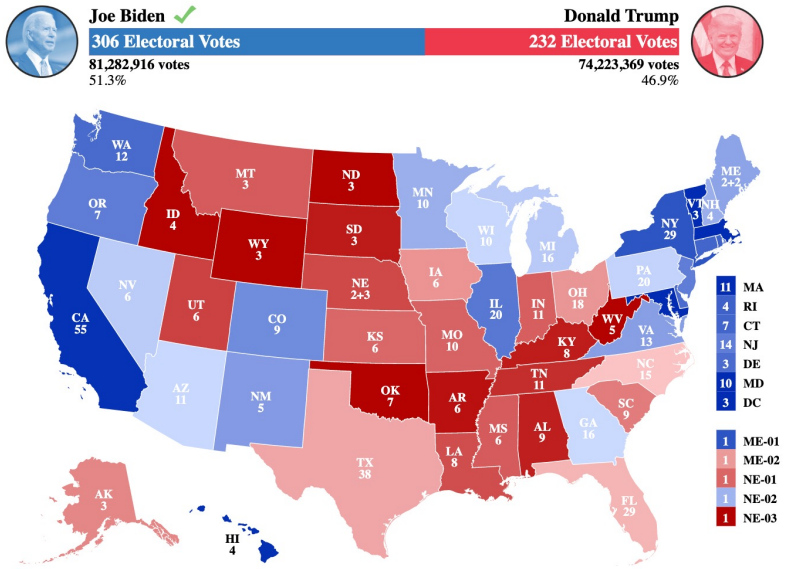
Values of a statistic vary because random samples vary

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

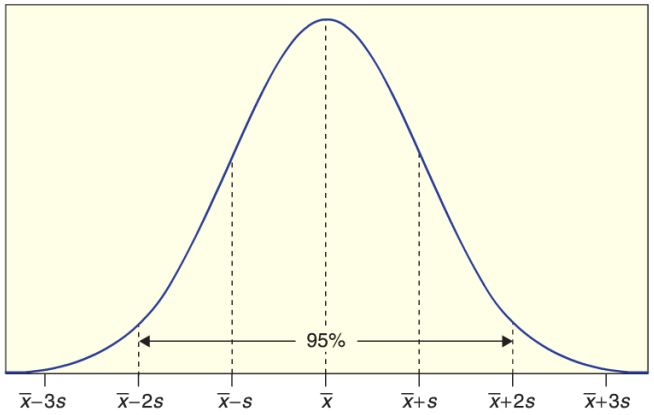
π_{Biden}

n = 1,000



$\hat{p}_{Biden}$

$\hat{p}_{Biden}$

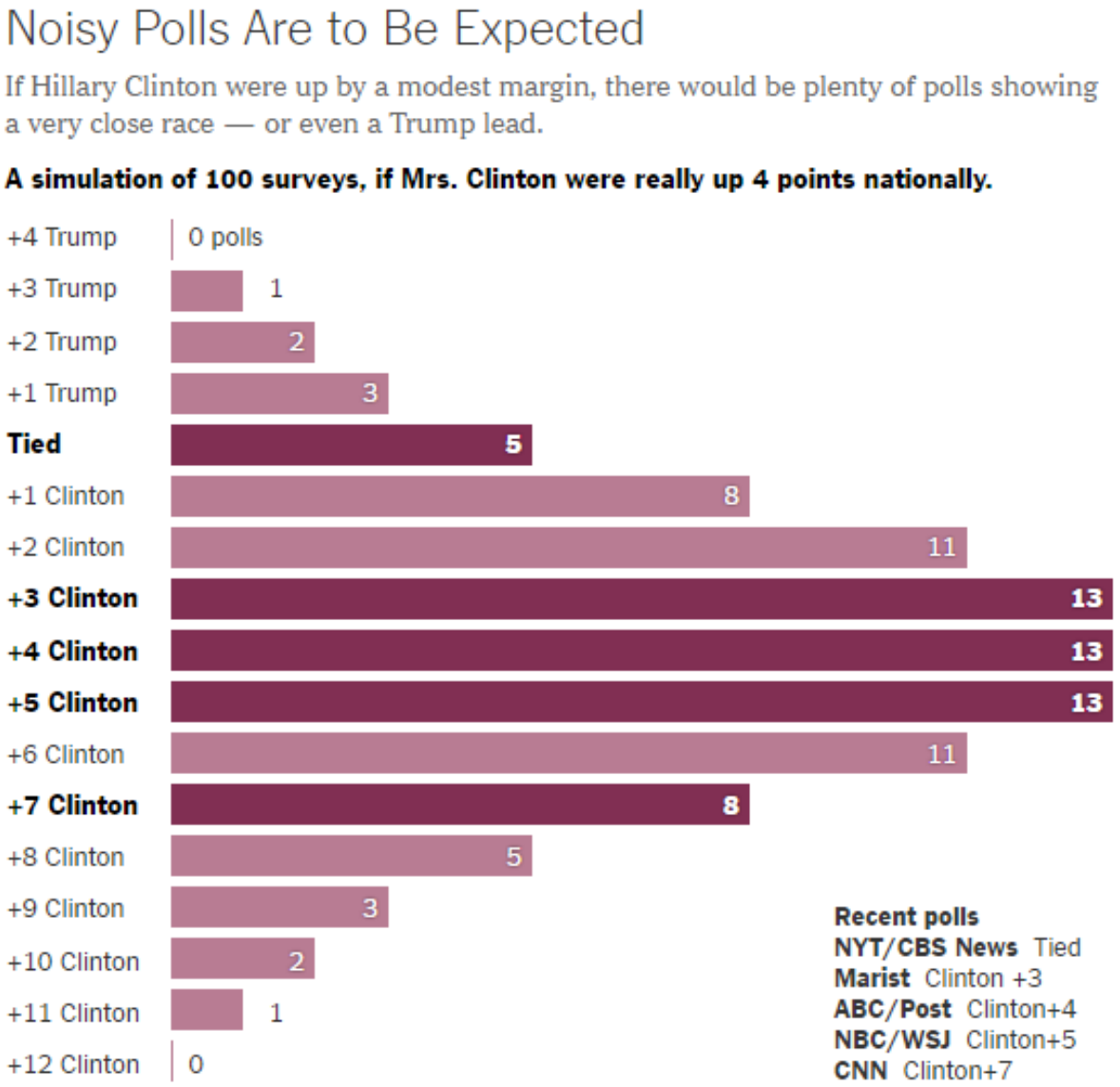


$\hat{p}_{Biden}$

Sampling distribution!

Let's explore this in Jupyter!

Confused by Contradictory Polls? Take a Step Back



What is this called?

What parameter are they trying to estimate?

NY Times 7/18/16

Hypothesis tests

A quick note on probability

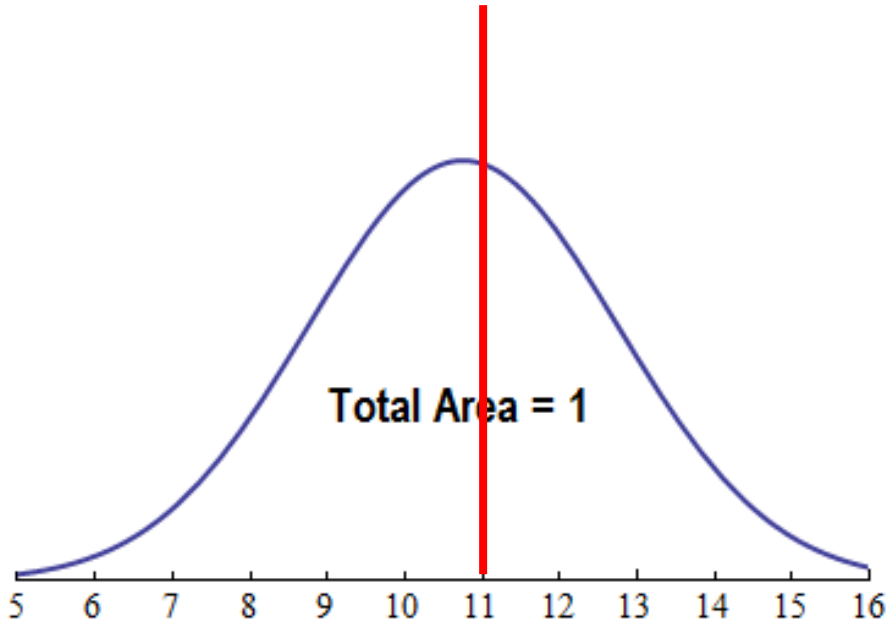
Probability is a way of measuring the likelihood that an event will occur.

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.

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.

Example 2: Trump Has Slight Lead Over DeSantis In GOP Primary, Quinnipiac University National Poll Finds

How can we write this using symbols?

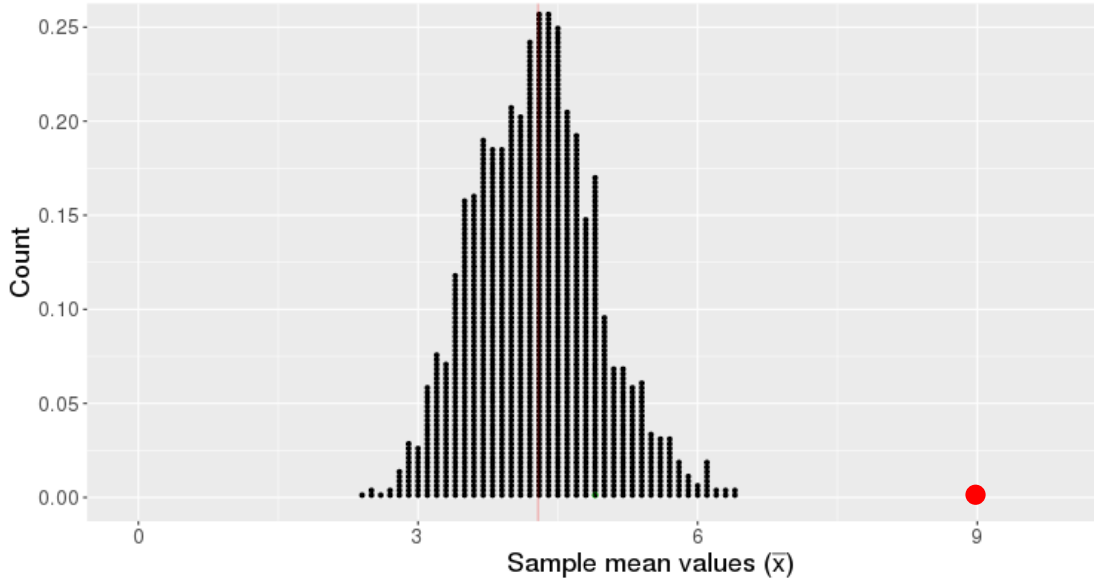
- $\pi_{Trump} > \pi_{DeSantis}$ or $\pi_{Trump} - \pi_{DeSantis} > 0$

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.

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. The 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 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 out 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

Motivating example: The Bechdel 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 more 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: $\hat{p} = 803/1794 = 0.448$

Chance models

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 = passed the Bechdel test

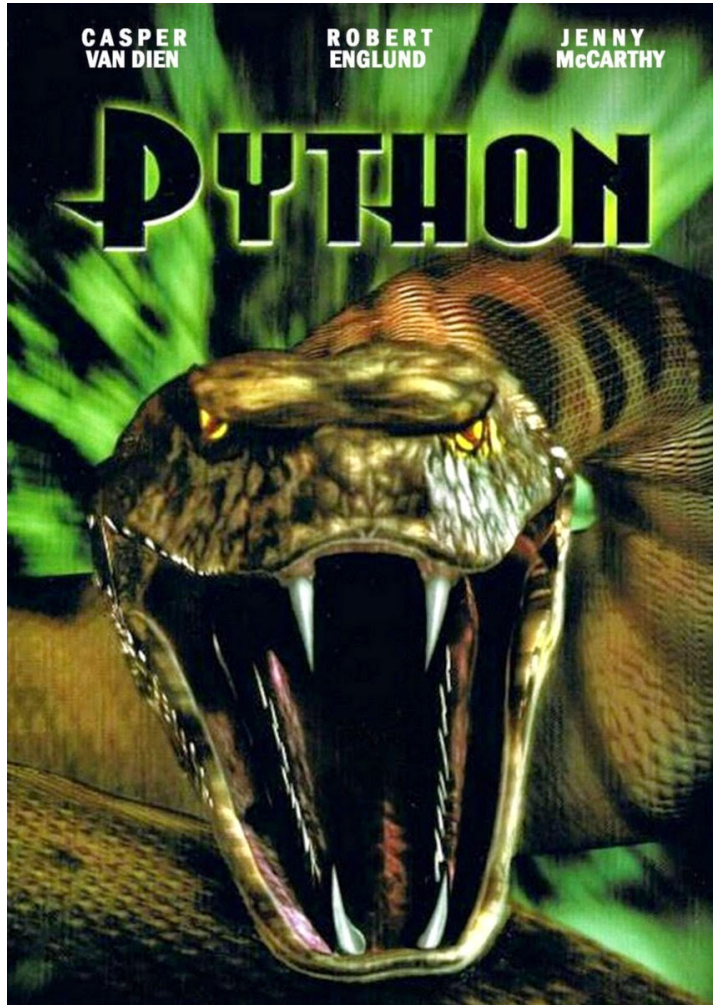
Tails = failed to pass the Bechdel test

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

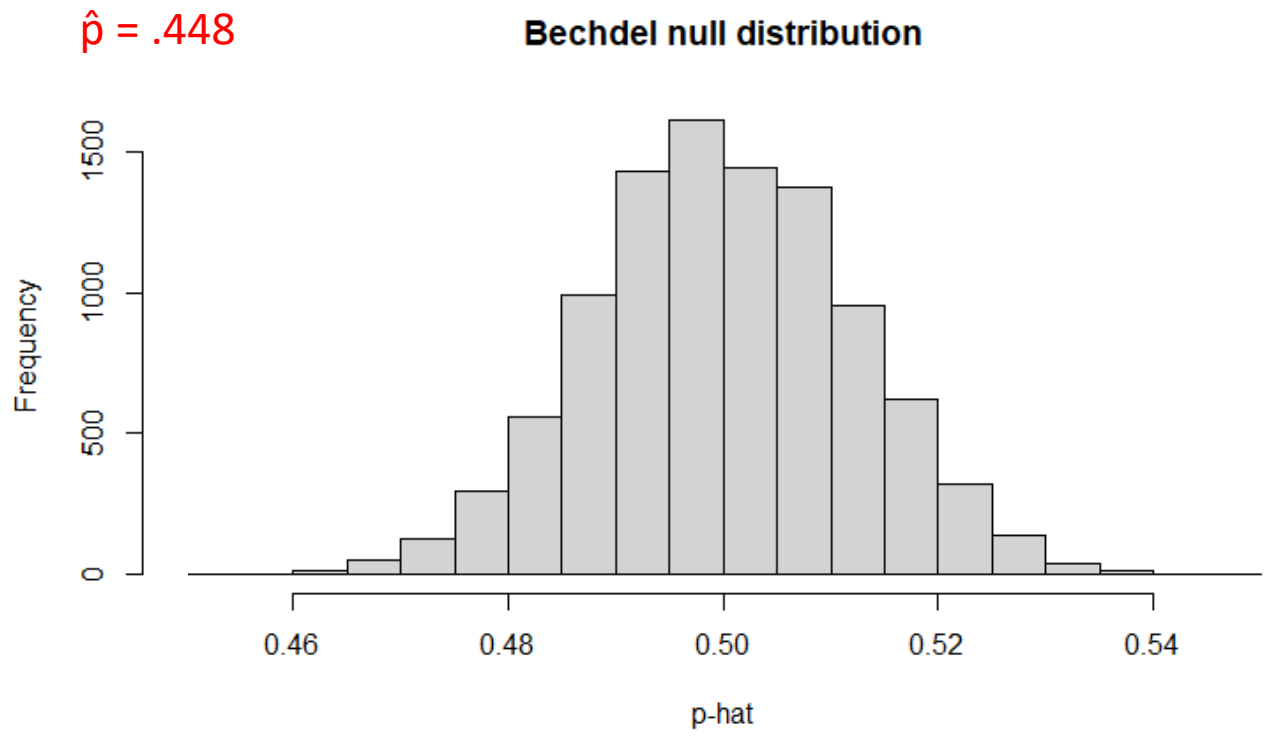
Chance models

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



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

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

Q: What can be conclude?

Let's try it in Python