

# Inference for Data Science

## Learning goals:

- Understand what a random sample is.
  - Appreciate the importance of random sampling.
  - Understand what a confidence interval states.
  - Learn common use cases for confidence intervals.
  - Compare big data and a random sample.
- 

**COGS 108 Fall 2019**  
**Lecture 10**  
**Sam Lau**

**samlau.me**  
**lau@ucsd.edu**  
**OH: Wed 10-11a in SSRB 100**

# Who is Sam?

**2nd year Ph.D. student in Cog Sci  
advised by Philip Guo**

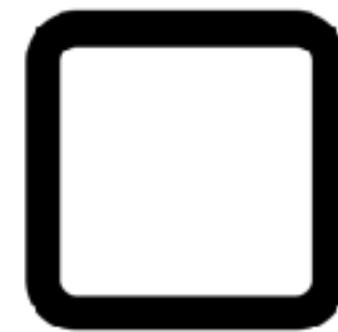
**Research: computational tools to  
teach data science**

**Previously taught data science @  
Berkeley**

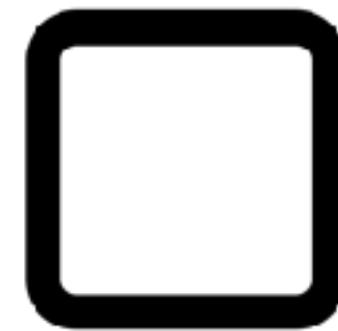
**(TA 5 times, Instructor 2 times)**

**Wrote a textbook for data science:  
[textbook.ds100.org](http://textbook.ds100.org)**

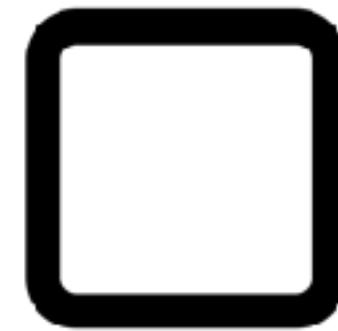




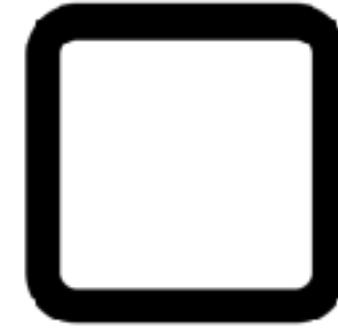
**Why random samples are important**



**What a random sample is**



**What a confidence interval is**



**Why confidence intervals are important**

# The Literary Digest

NEW YORK

OCTOBER 31, 1936

---

---

## *Topics of the day*

**LANDON, 1,293,669; ROOSEVELT, 972,897**

**Final Returns in The Digest's Poll of Ten Million Voters**

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

American National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens

Chicago Daily Tribune Home

# DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State: Boyle Leads in City

REPUBLICAN RECORD CITY  
TICKET LEADS VOTE SEEN IN  
LATE TALLIES

VALUING IN LATE

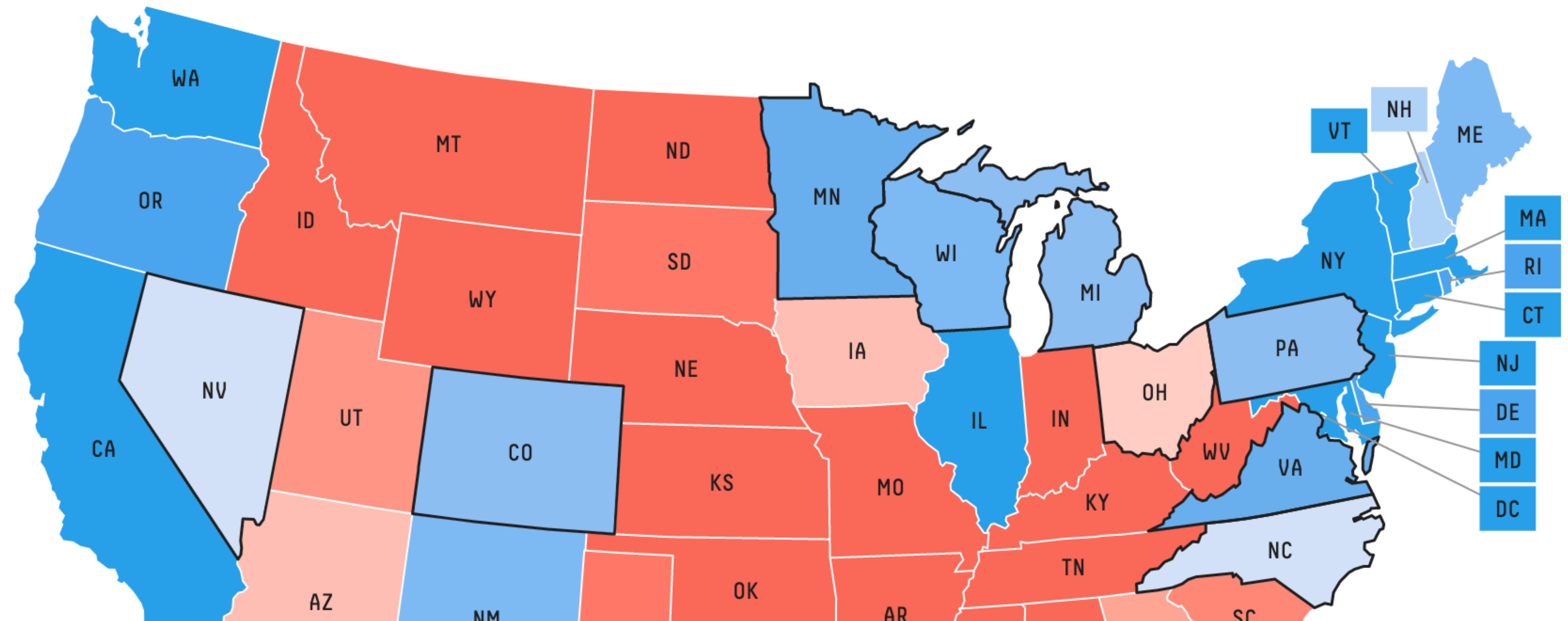
BEST CITIES  
CITIES  
TICKET LEADS  
VOTE SEEN IN  
LATE TALLIES





G  
E  
P

# Chance of winning



**What went wrong?**

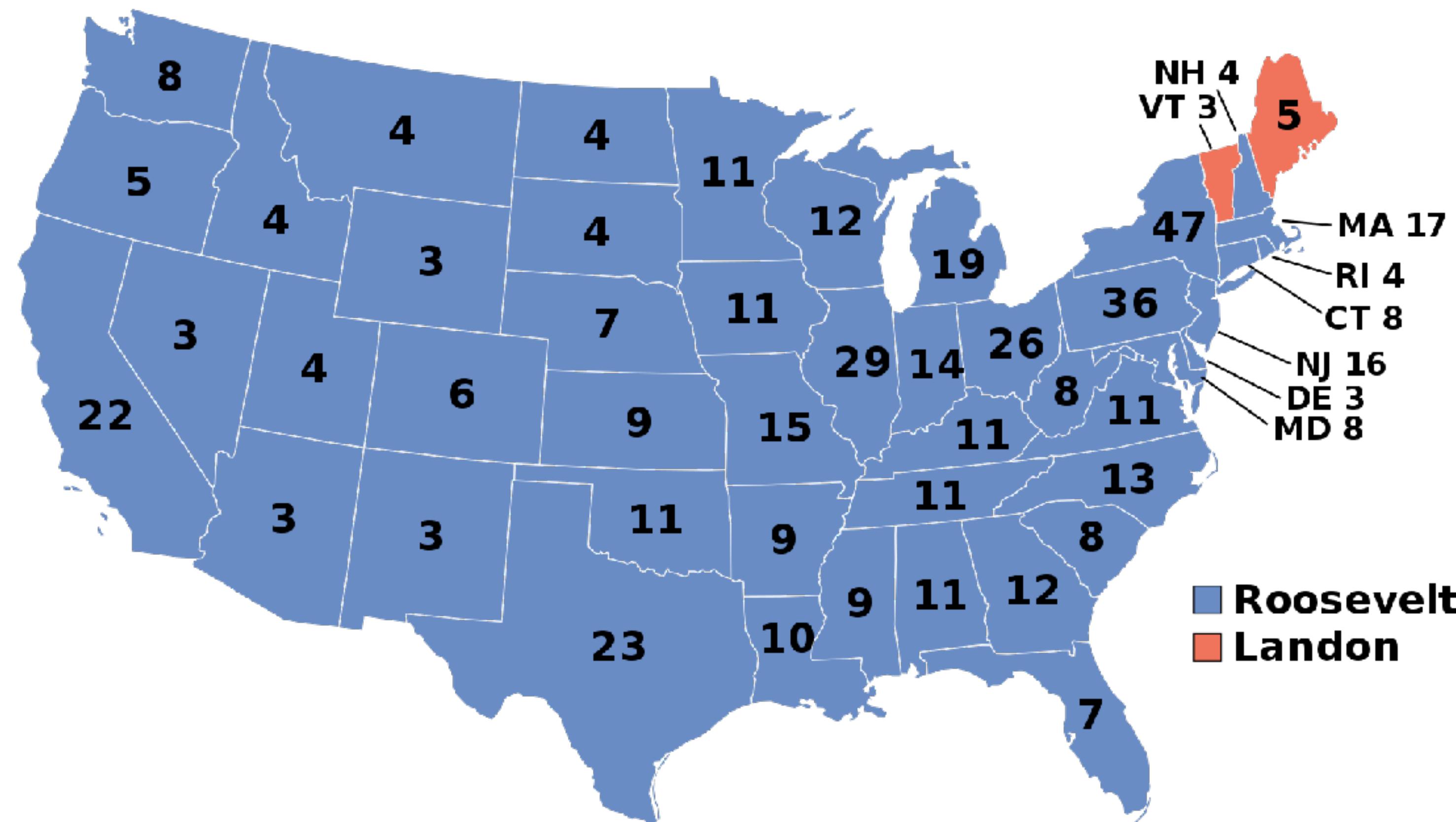
# 1936 Election

Literary Digest predicted results  
of last 5 elections.

Mailed 10 million  
questionnaires to readers.  
2.3 million responded.

Results off by 20%!  
Gallup Poll correct with just  
50,000 responses.

What happened?



# 1948 Election

Gallup Poll used quota sampling:  
interviewers had to match right  
distribution of sex, ethnicity, income  
level, age, etc.

Sampled 3,250 people.

Results off by 10%

**What happened?**



# 2016 Election

538 used a combination of many local polls with an emphasis on likely voters.

>1000 polls with over 1 million likely voters. Special adjustments for bias in individual polls.

Predicted 30% chance of Trump win.

**What happened?**

The Polls Missed Trump. We Asked Pollsters Why.

By [Carl Bialik](#) and [Harry Enten](#)  
Filed under [2016 Election](#)

The polls missed Donald Trump's election. Individual polls missed, at the state level and nationally (though national polls weren't [far off](#)). So did aggregated polls. So did [poll-based forecasts](#) such as [ours](#). And so did [exit polls](#).

The miss wasn't unprecedented or even, these days, all that unusual. Polls [have missed](#) recent elections in the U.S. and abroad by margins at least as big. Every poll, and every prediction based on it, is probabilistic in nature:

Making good predictions requires a  
**random sample** from the  
**population of interest.**

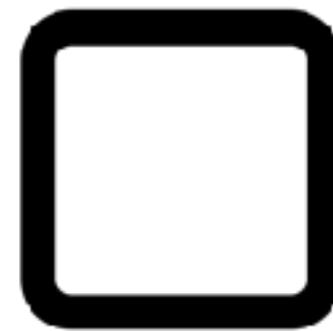
(Or, collect your entire population in a **census**.)

Non-random samples will be  
**biased** in unpredictable ways.

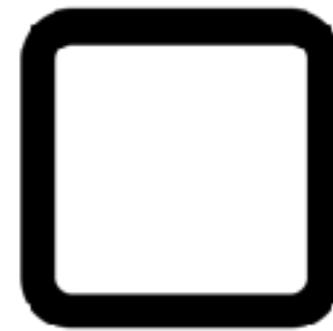
Data science and statistics  
**cannot fix** this intrinsic bias.



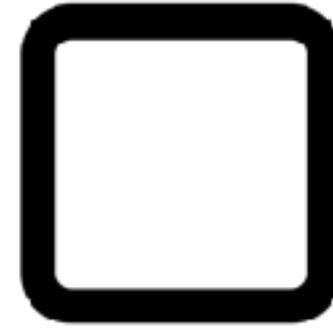
**Why random samples are important**



**What a random sample is**



**What a confidence interval is**



**Why confidence intervals are important**

# Types of Samples

Suppose I want to know the average height of COGS108 students.

**Population of interest:** Heights of COGS108 students.

**Census:** Measuring height of every student in 108.

**Non-Random Sample:** Picking my favorite 20 students and measuring their heights.

**Random Sample:** Mixing all student names in a hat, drawing 20 students, then measuring those 20.

# Random sample:

Every individual in the population  
has an **exact probability** of  
appearing in the sample.

# What about my final project dataset?

Dataset	Pop'n of Interest	Verdict
US Census income report	US citizens	Census
US Census income report	California residents	Census
All traffic accidents in San Diego from 2015-2019	Traffic accidents in SD in 2018	Census
All traffic accidents in San Diego from 2015-2019	Future traffic accidents in SD	Non-random sample

# What about my final project dataset?

Dataset	Pop'n of Interest	Verdict
Top 100 songs on Spotify in 2018	Popular Spotify songs	Non-random sample
Top 100 songs on Spotify in 2018	Popular Spotify songs in 2018	Non-random sample
All daily trending YouTube videos in 2019	Trending YouTube videos in Sept 2019	Census
All daily trending YouTube videos in 2019	Future trending YouTube videos	Non-random sample

# How to be less wrong

Be **precise** about your population and your sample.

If you have a census, **don't use p-values** or confidence intervals.

If you have a non-random sample:

**Explicitly say so**, outline potential sources of bias, and anticipate how bias might affect your results.

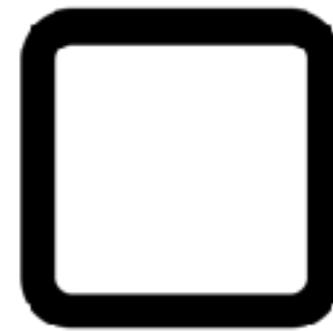
“In this analysis, we assume that our dataset of YouTube videos this year is representative of YouTube videos next year as well. This can cause the following problems with our analysis...”



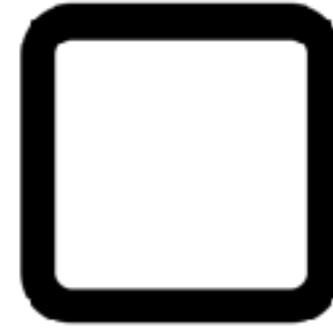
**Why random samples are important**



**What a random sample is**



**What a confidence interval is**



**Why confidence intervals are important**

**5 min break**

# **What is inference?**

# Probability tells you about the sample

## Population

48% voted for Clinton

46% voted for Trump

6% voted for other



## Sample of 100 voters

What's the probability  
that sample will have  
**≥51** people who voted  
for Clinton?

# Inference tells you about the population

**Population**

**??%** voted for Clinton

**Sample of 100 voters**

51 people in sample  
voted for Clinton. What  
percent of population  
voted for Clinton?



# We can estimate single values...

**Population**

**51%** voted for Clinton



**Sample of 100 voters**

51 people in sample  
voted for Clinton. What  
percent of population  
voted for Clinton?

# ...but estimating ranges is more useful.

## Population

**51%  $\pm$  2%** voted for Clinton.

Or: **between 49% and 53%** voted for Clinton.



## Sample of 100 voters

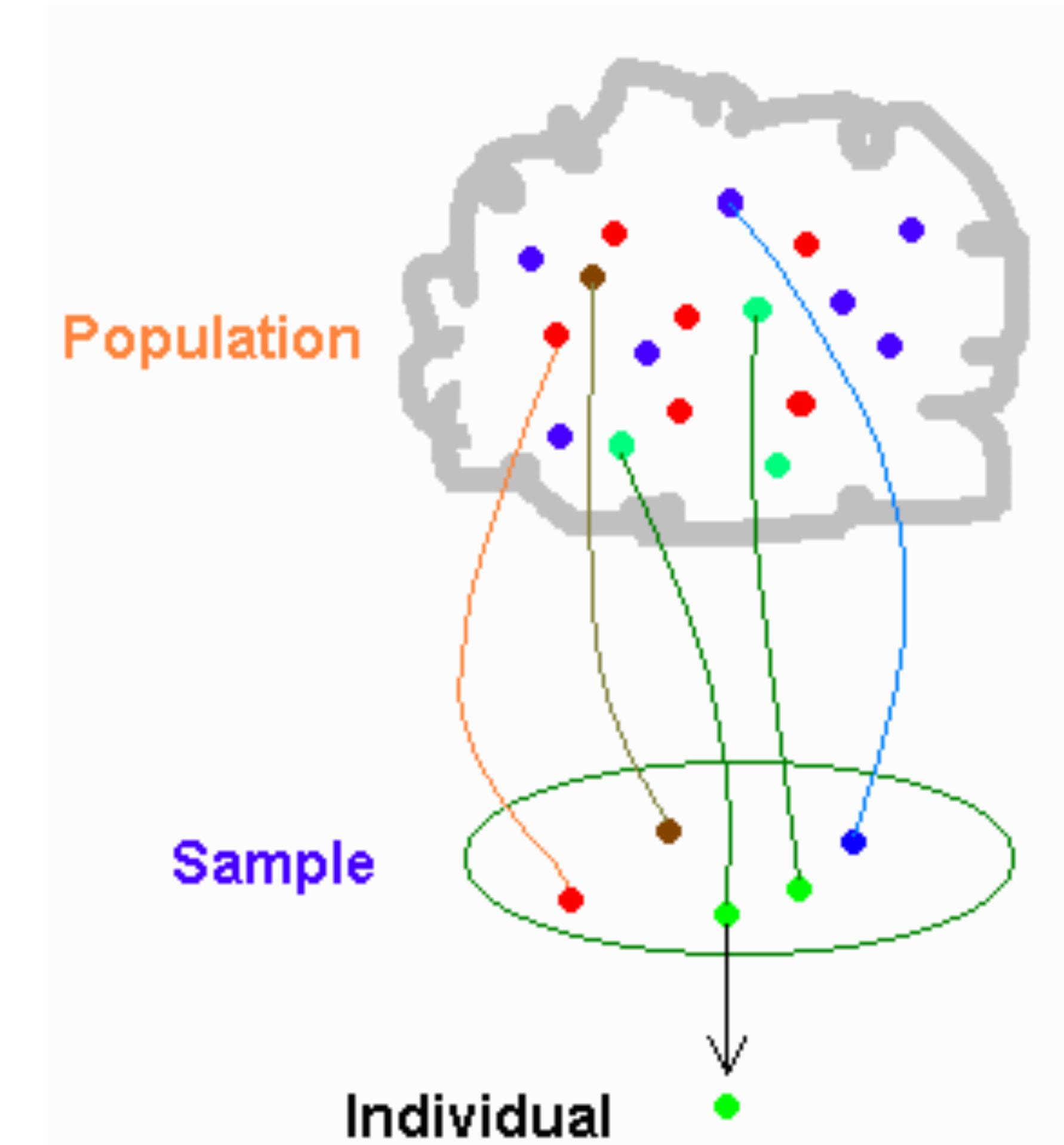
51 people in sample voted for Clinton. What percent of population voted for Clinton?

# Prediction error comes from:

**Sampling bias**

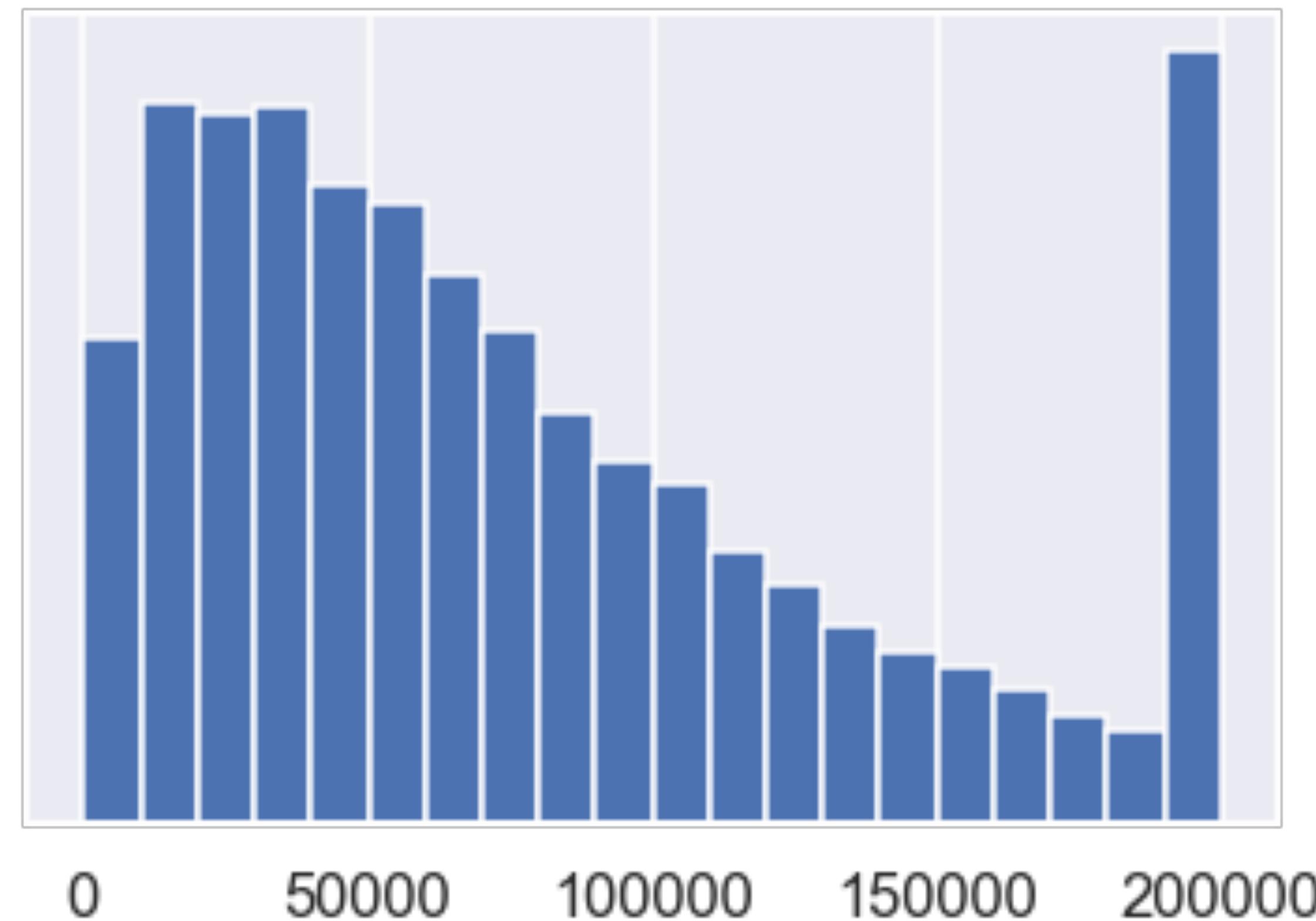


**Sampling variability**

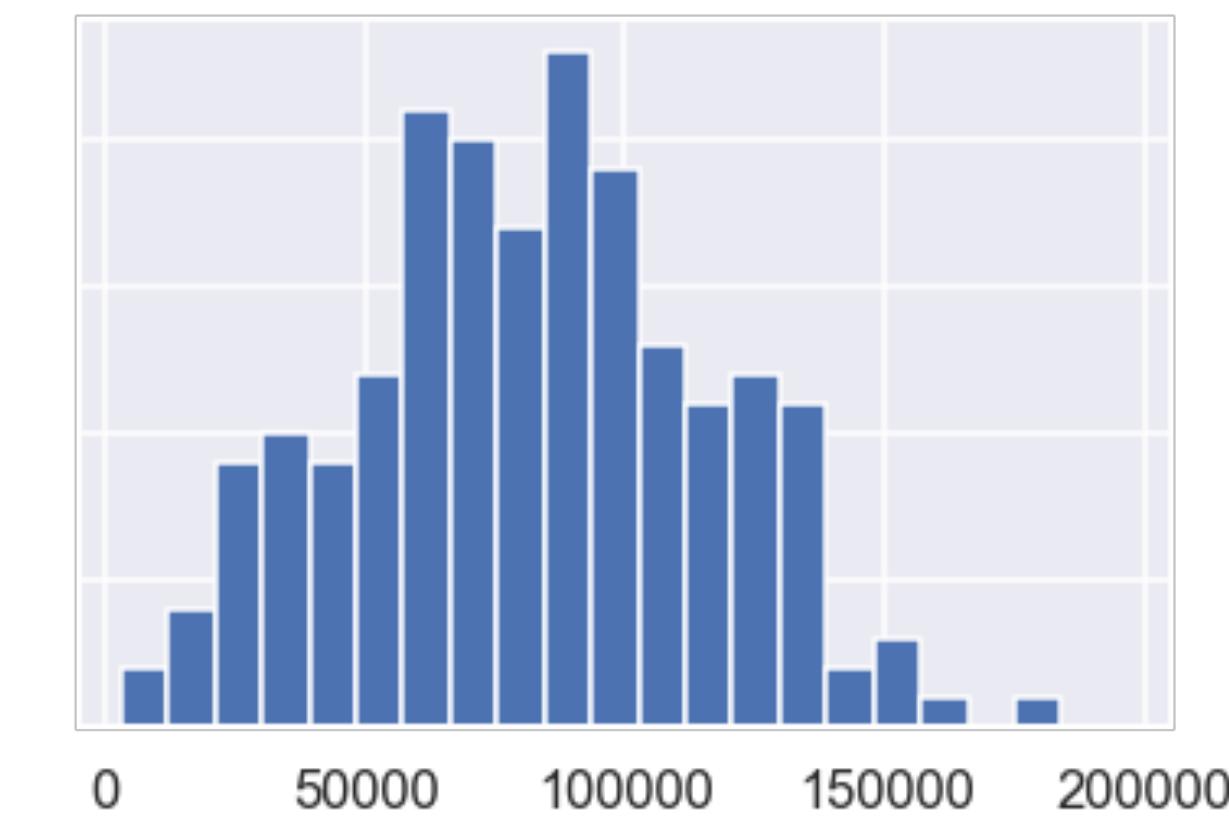


# Prediction error comes from:

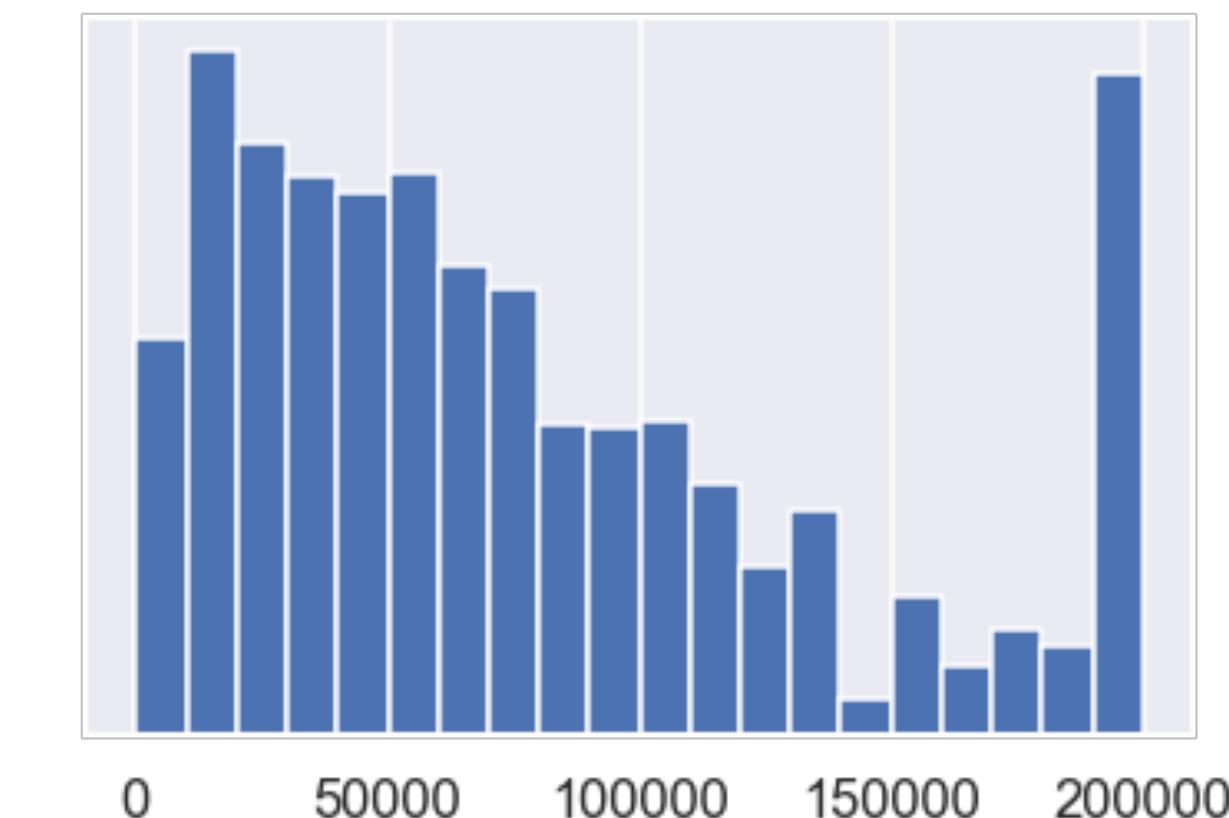
Household Incomes in the US



**Sampling bias**



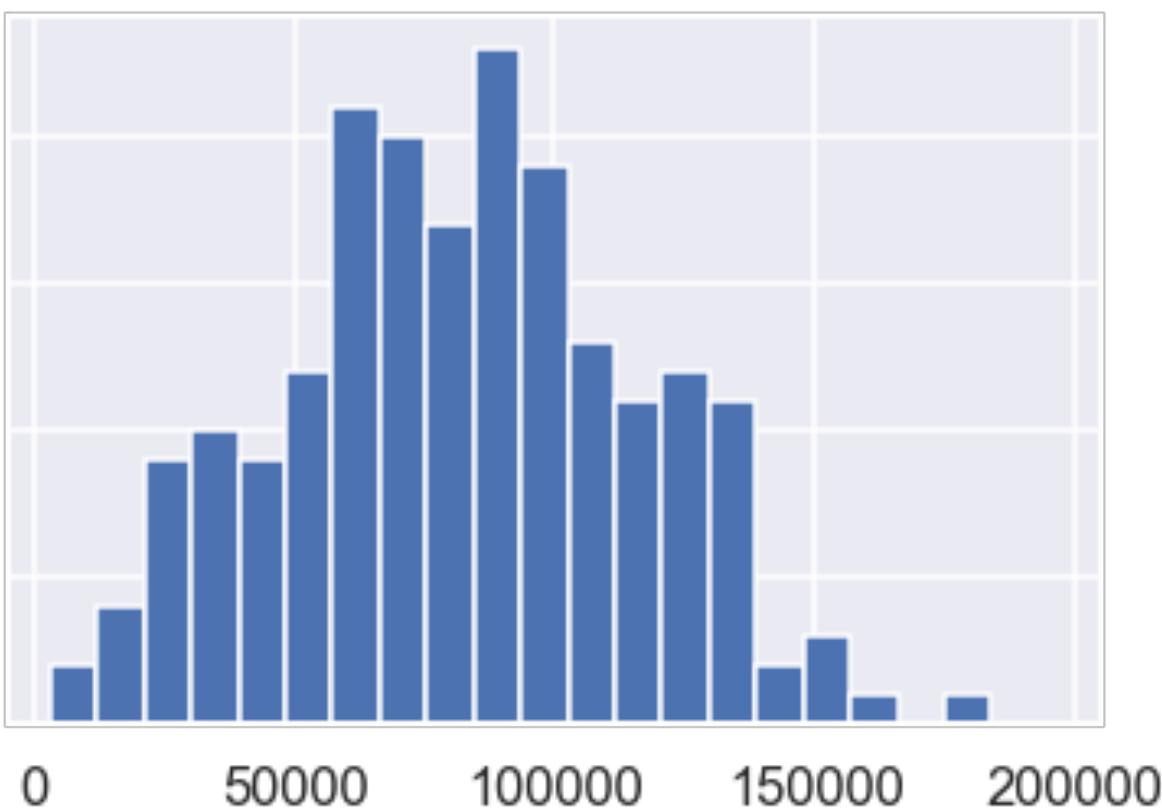
**Sampling variability**



<https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-01.html>

# Inference only accounts for sampling variability

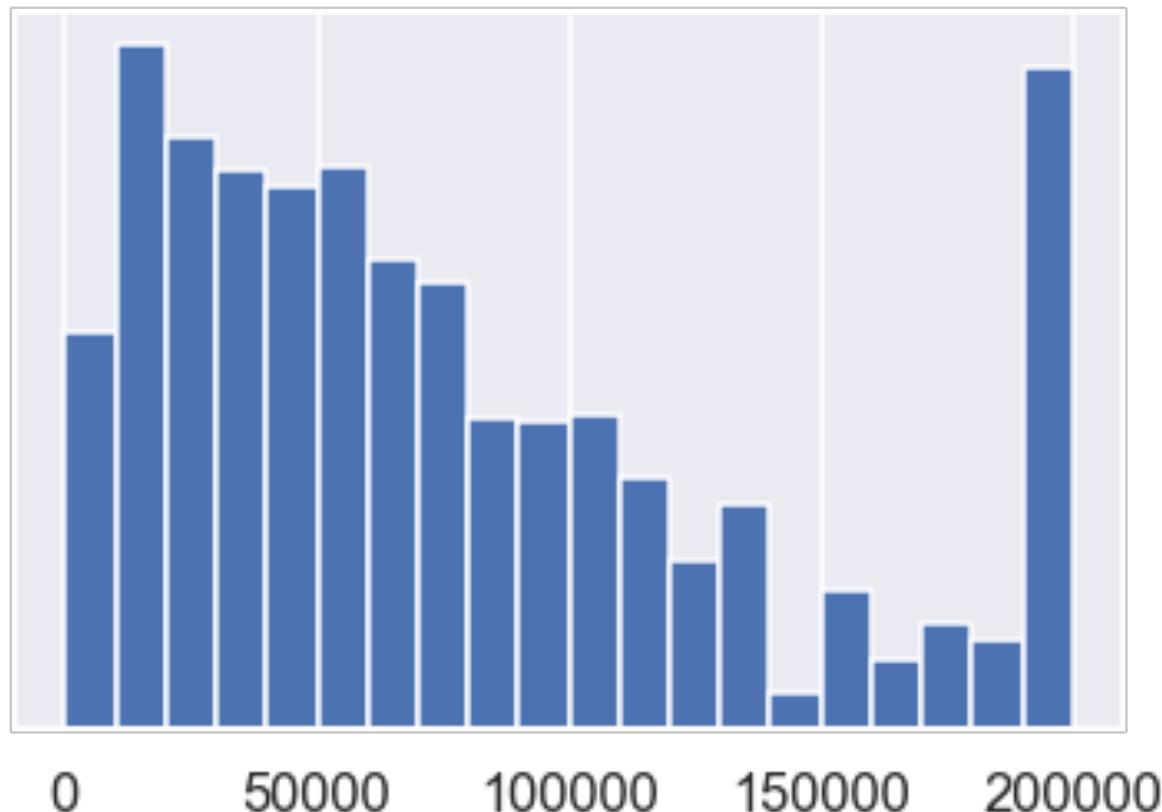
## Sampling bias



You need to re-examine your data collection strategy.

**Statistics will not help you** (much).

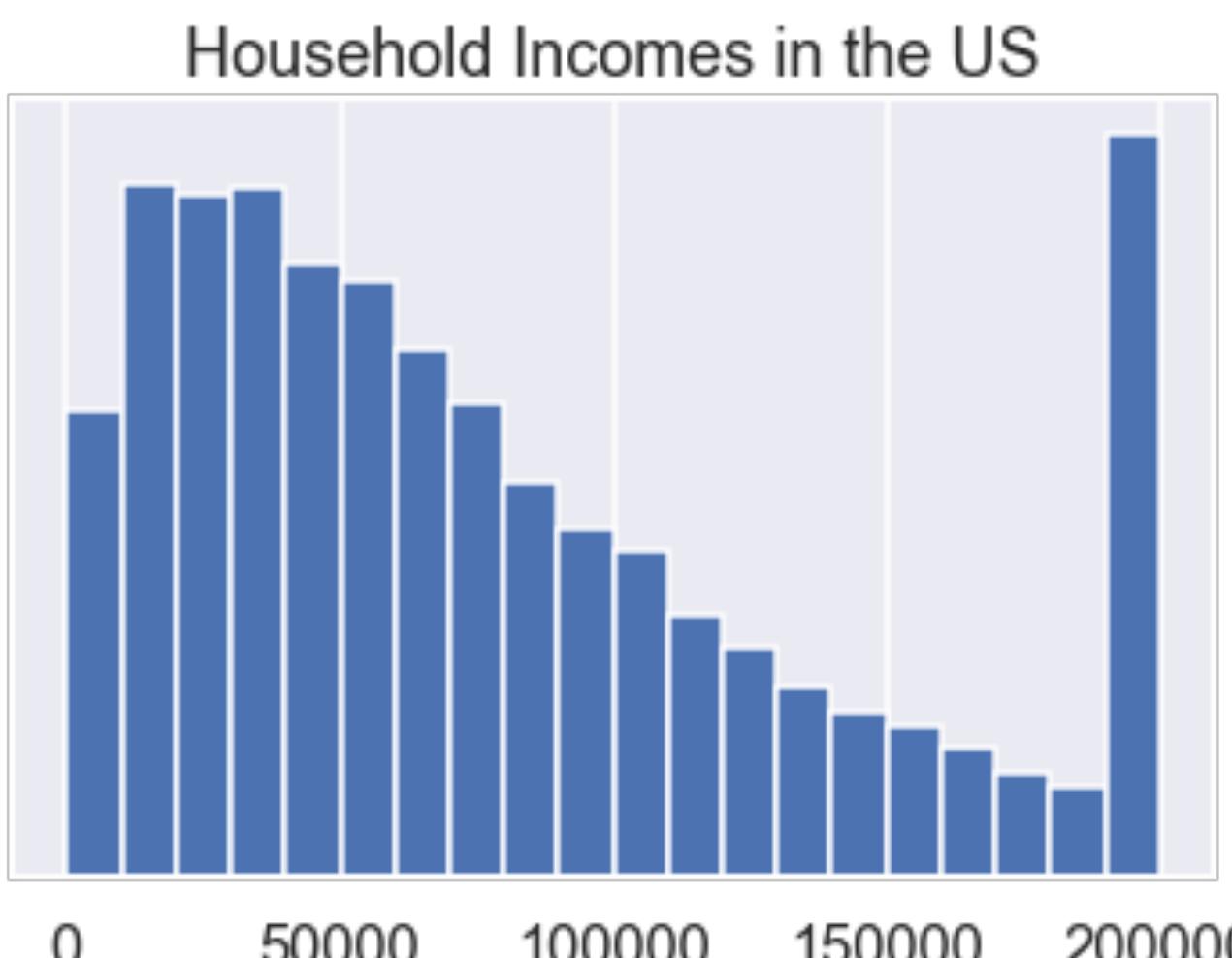
## Sampling variability



You can account for this uncertainty using inference techniques like **confidence intervals**.

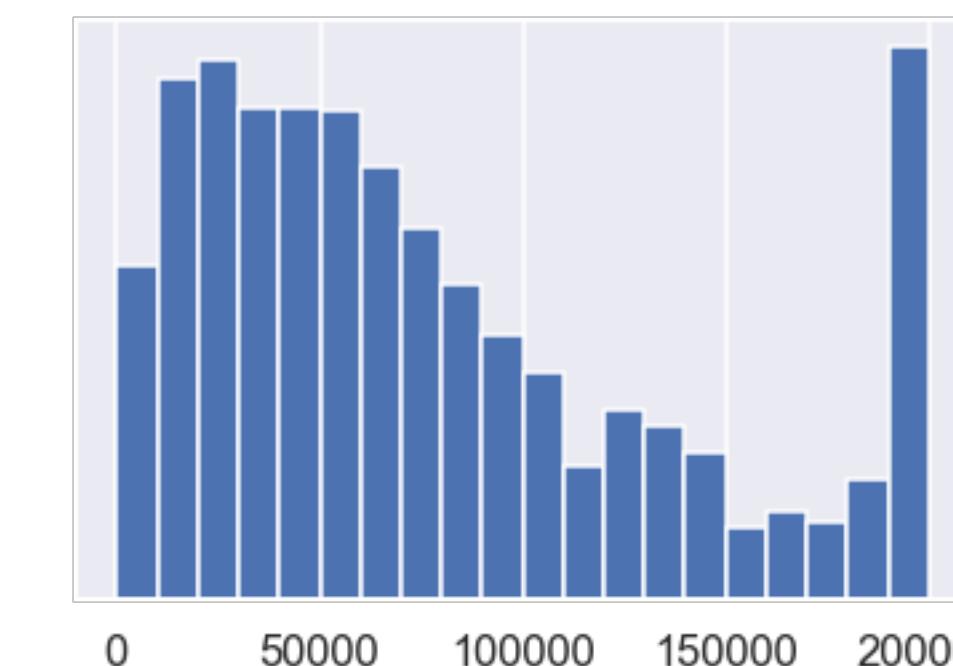
# Making Confidence Intervals

## Population



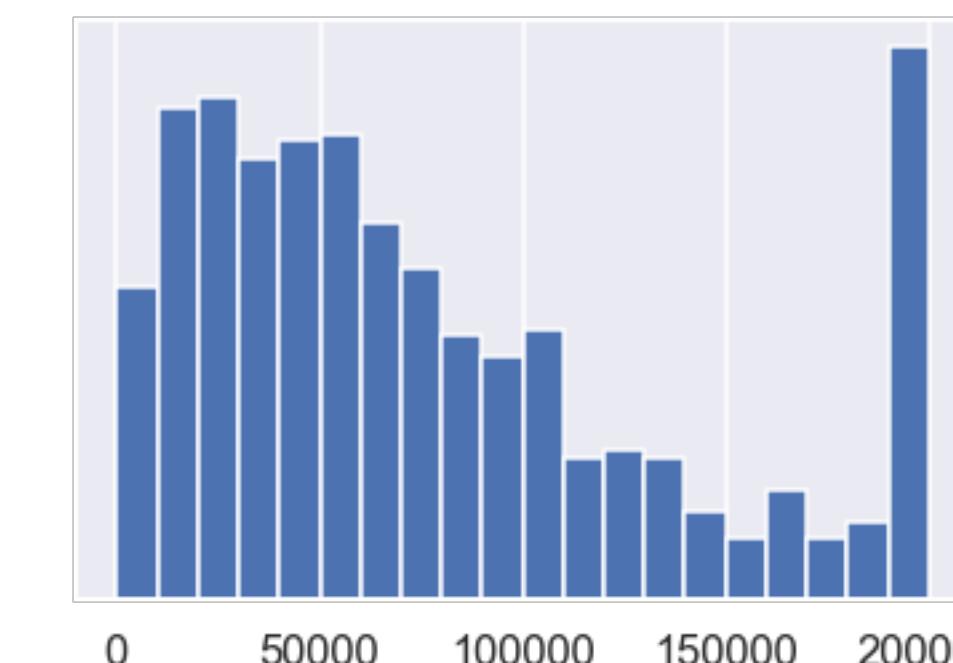
Mean: \$63,000

## Samples



## Sample means

\$62,000



\$61,000



\$65,000

# Making Confidence Intervals

## Sample means

\$62,000

\$61,000

\$65,000

\$59,000

\$61,000

\$60,000

\$63,000

\$61,000

\$65,000

\$65,000

## Take the middle 95%

The 95% confidence interval for mean population income is (\$60000, \$65000).

Collecting many samples is  
expensive...

...so we use statistical techniques  
to make CIs using a **single sample**.

# In Practice

$$\boxed{\bar{X}} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$

**My CI is centered at my sample mean.**

# In Practice

$$\bar{X} \pm [t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}]$$

My CI is centered at my sample mean.

**If I want more confidence, I need to make my interval wider.**

# In Practice

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$

My CI is centered at my sample mean.

If I want more confidence, I need to make my interval wider.

**If my sample is large, my CI will be more narrow.**

# In Practice

$$\bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

My CI is centered at my sample mean.

If I want more confidence, I need to make my interval wider.

If my sample is large, my CI will be more narrow.

**Many other CI formulas follow this basic pattern.**

**If you like it...**

# ...put an interval on it.

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \frac{s}{\sqrt{s_{xx}}}$$
$$\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

# Takeaways

Confidence intervals quantify uncertainty from sampling.

(They do not fix broken samples.)

There's a confidence interval formula for almost any statistic you can imagine: mean, median, slope of regression line, etc.

One useful technique: the bootstrap lets you make confidence intervals for many types of statistics (learn more in DSC 10).



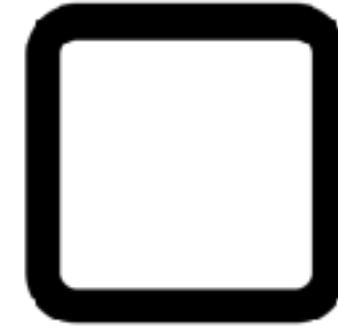
**Why random samples are important**



**What a random sample is**



**What a confidence interval is**



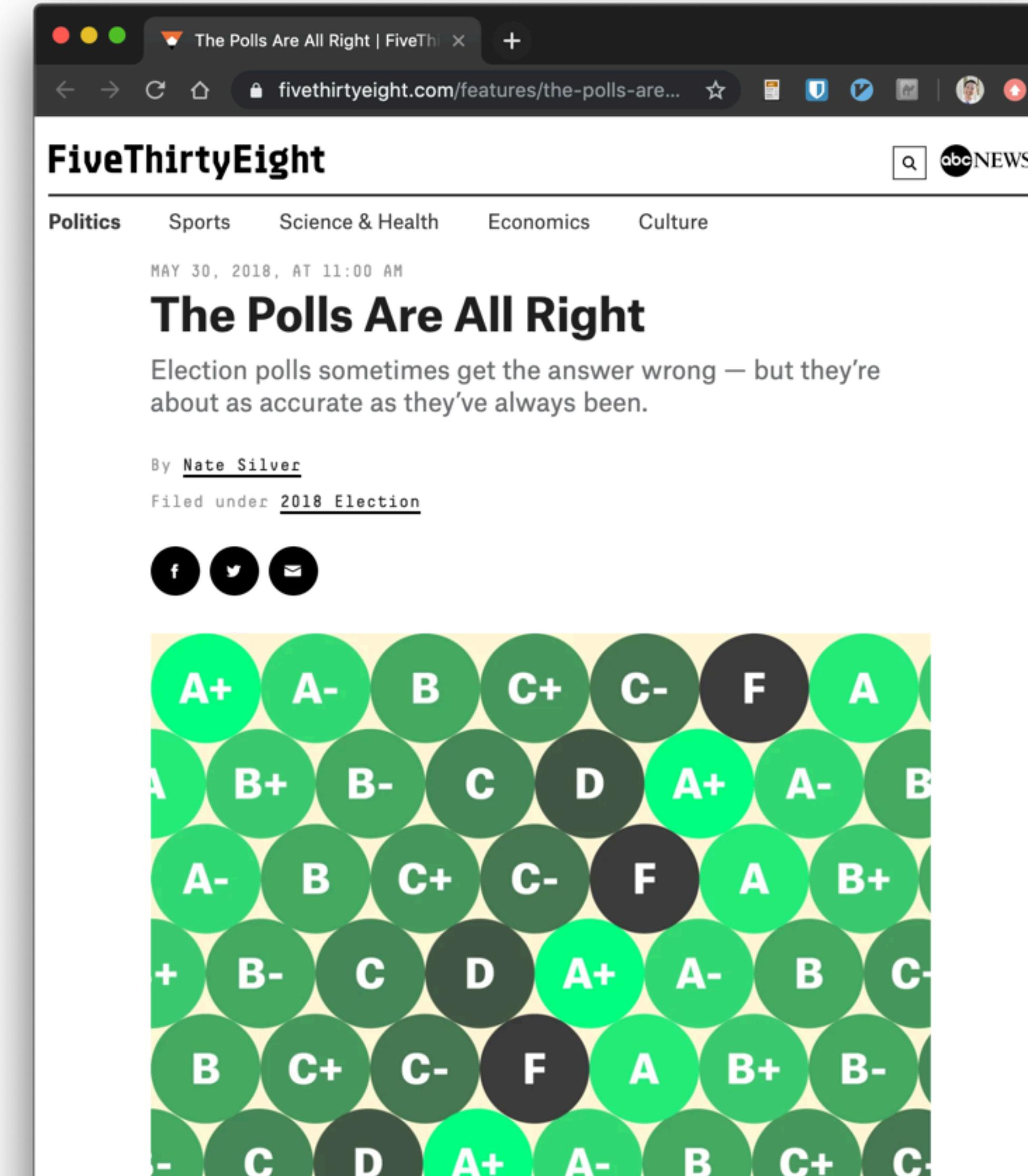
**Why confidence intervals are important**

# CIs let you account for sampling uncertainty

538 was wrong for 2016 election.

But error fell within historical polling errors.

CIs were actually accurate for previous elections.

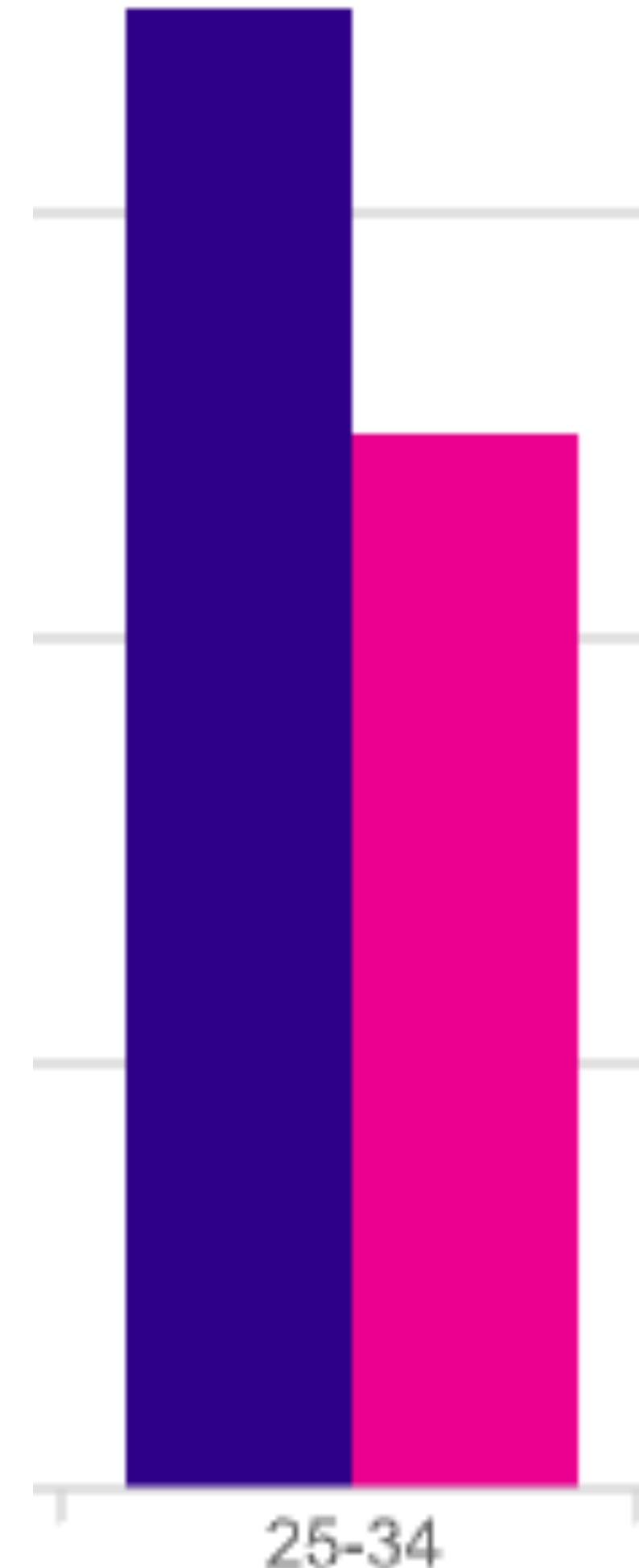


# CIs let you compare two quantities

Do more males smoke than females?

Make a CI for the **difference in means**.

If that CI **does not contain 0**, then you have evidence for a difference.

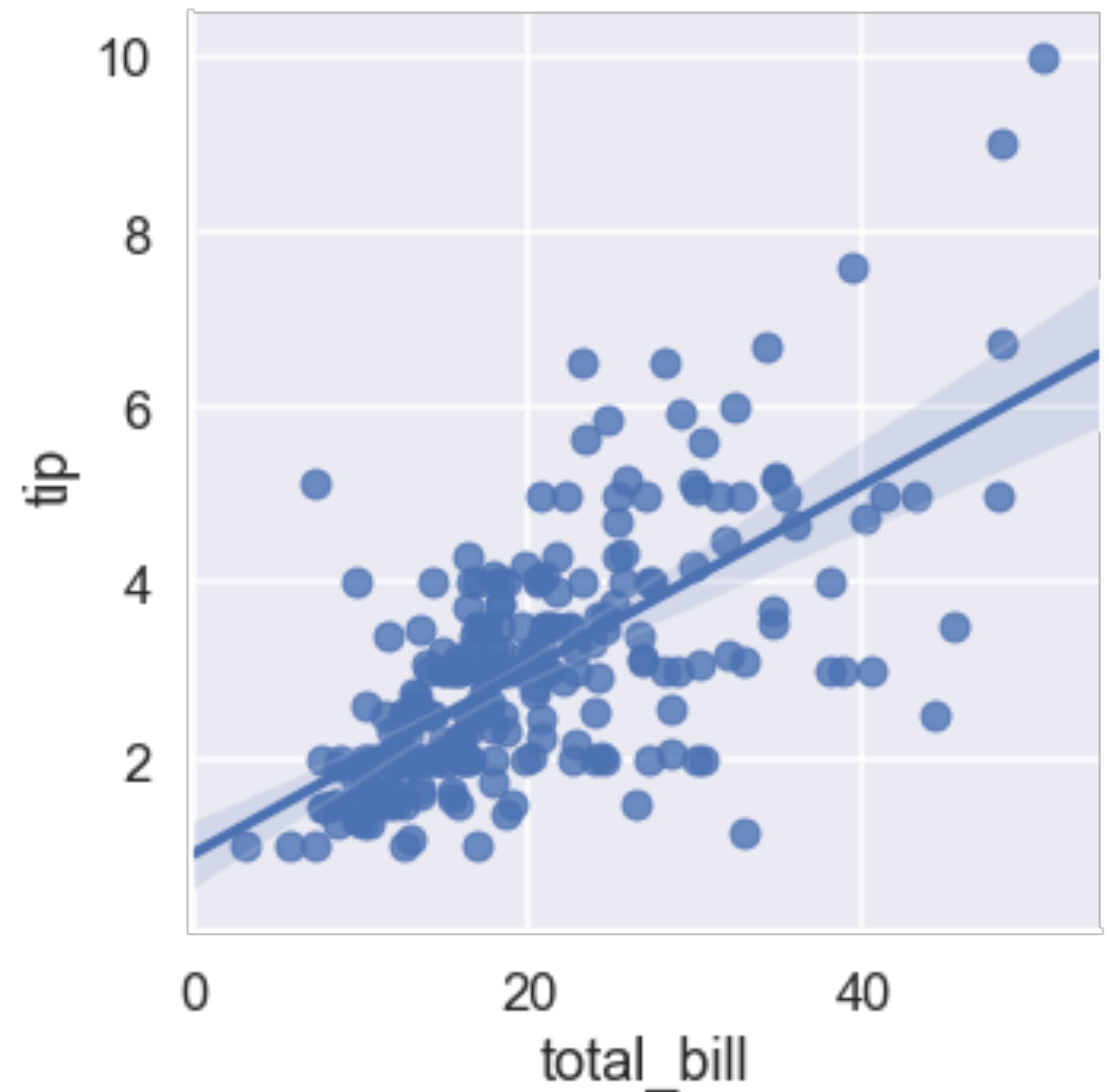


# CIs let you examine associations

Do people tip more if they buy more food?

Make a CI for the **slope of the regression line.**

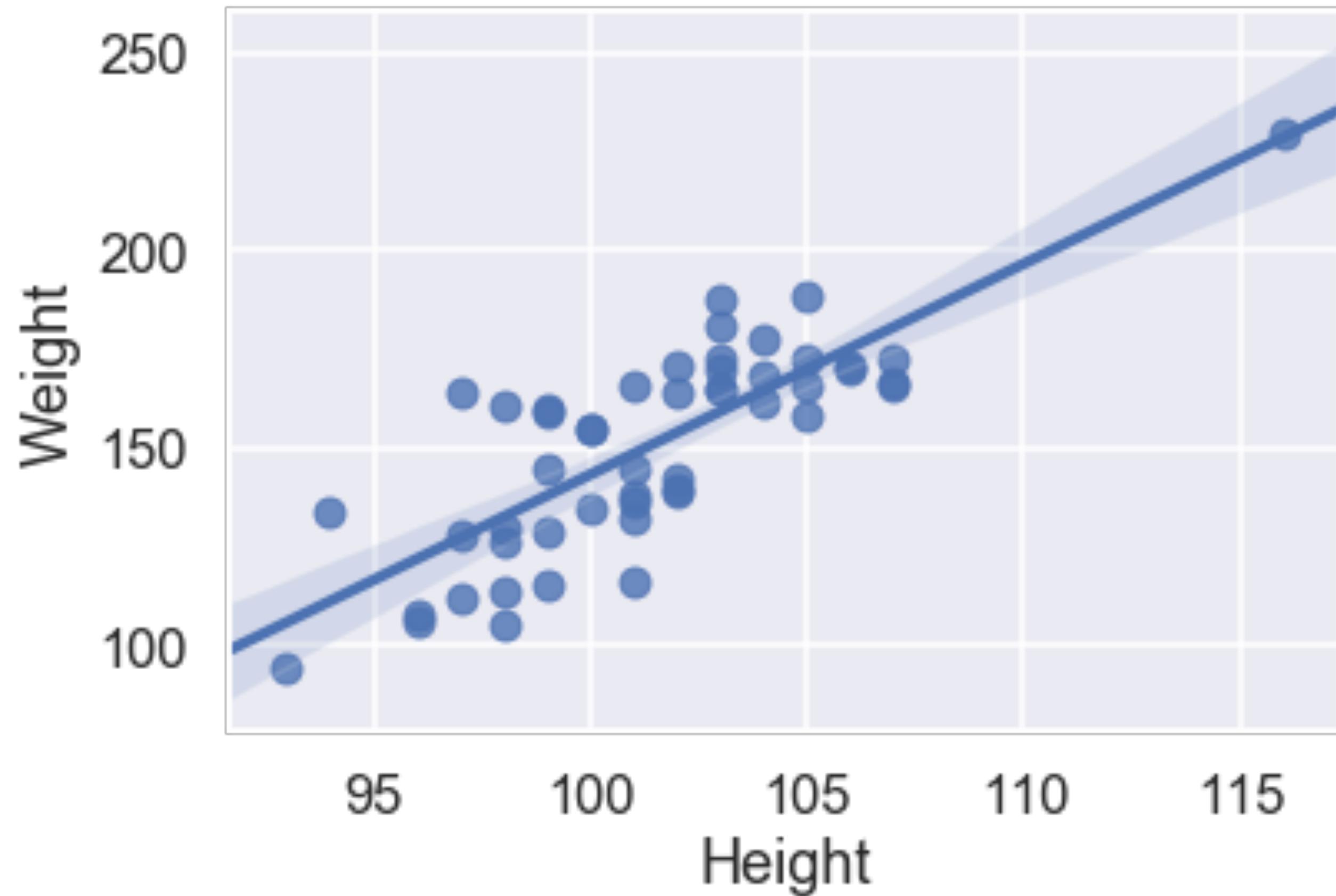
If the slope does not contain 0, you have evidence for an association.



# CIs let you examine predictions

How heavy will a donkey be if she has a height of 110 inches?

Make a CI for the **prediction of a regression model.**

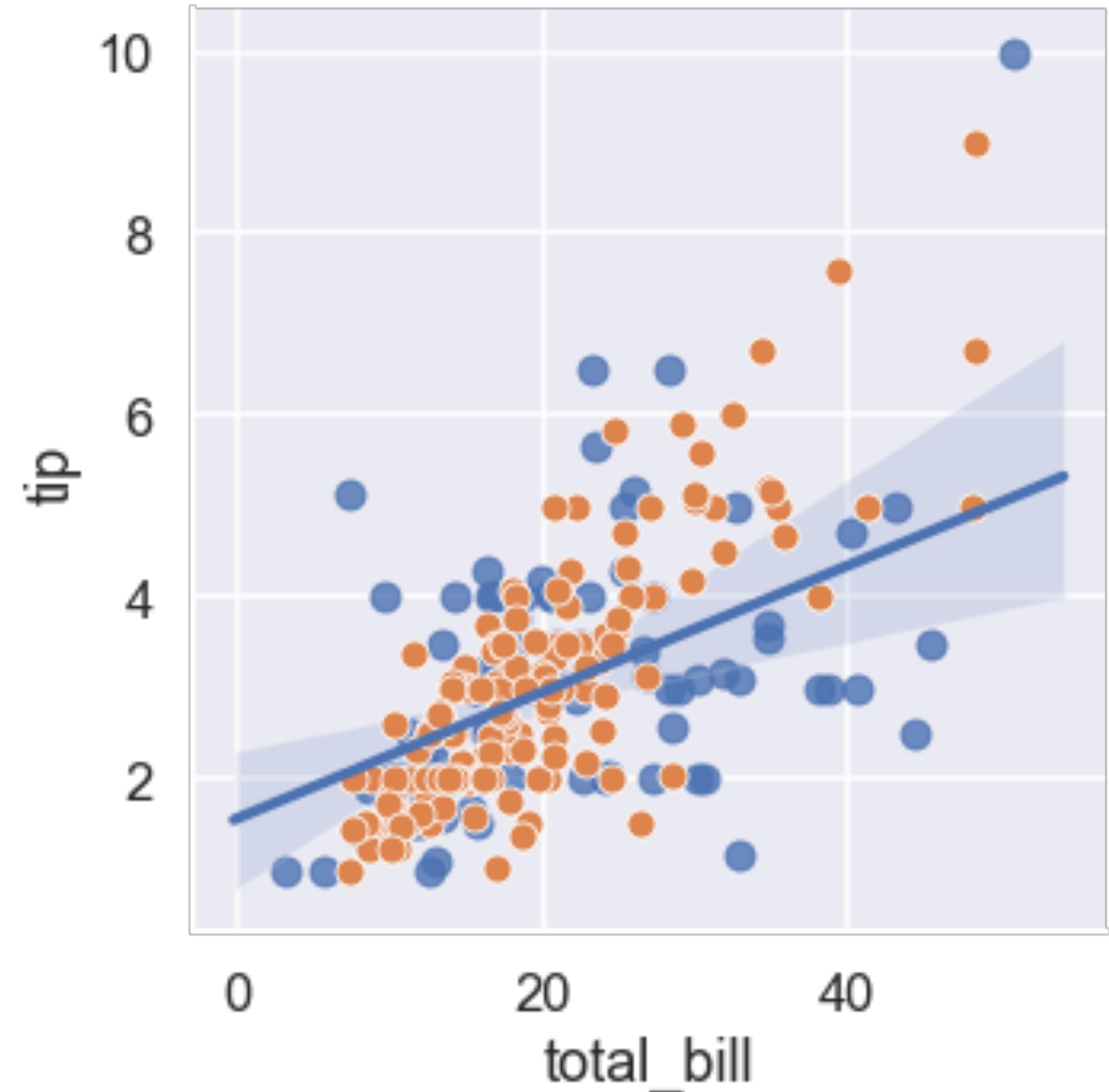


# CIs let you examine error

How well will your model perform on unseen data?

Make a CI for the error of your model **on your test set.**

This lets you estimate the **population error** of your model.





**Why random samples are important**



**What a random sample is**



**What a confidence interval is**



**Why confidence intervals are important**

**Can big data overcome the  
need for a random sample?**

# What would you trust more?

Random sample of size 400 (0.3% of U.S. voters)

Non-random sample of size 60,000,000 (50% of U.S. voters)

# Keep your hands up for SRS

Random sample of size 400

Non-random sample of 80% of U.S. voters

Non-random sample of 90% of U.S. voters

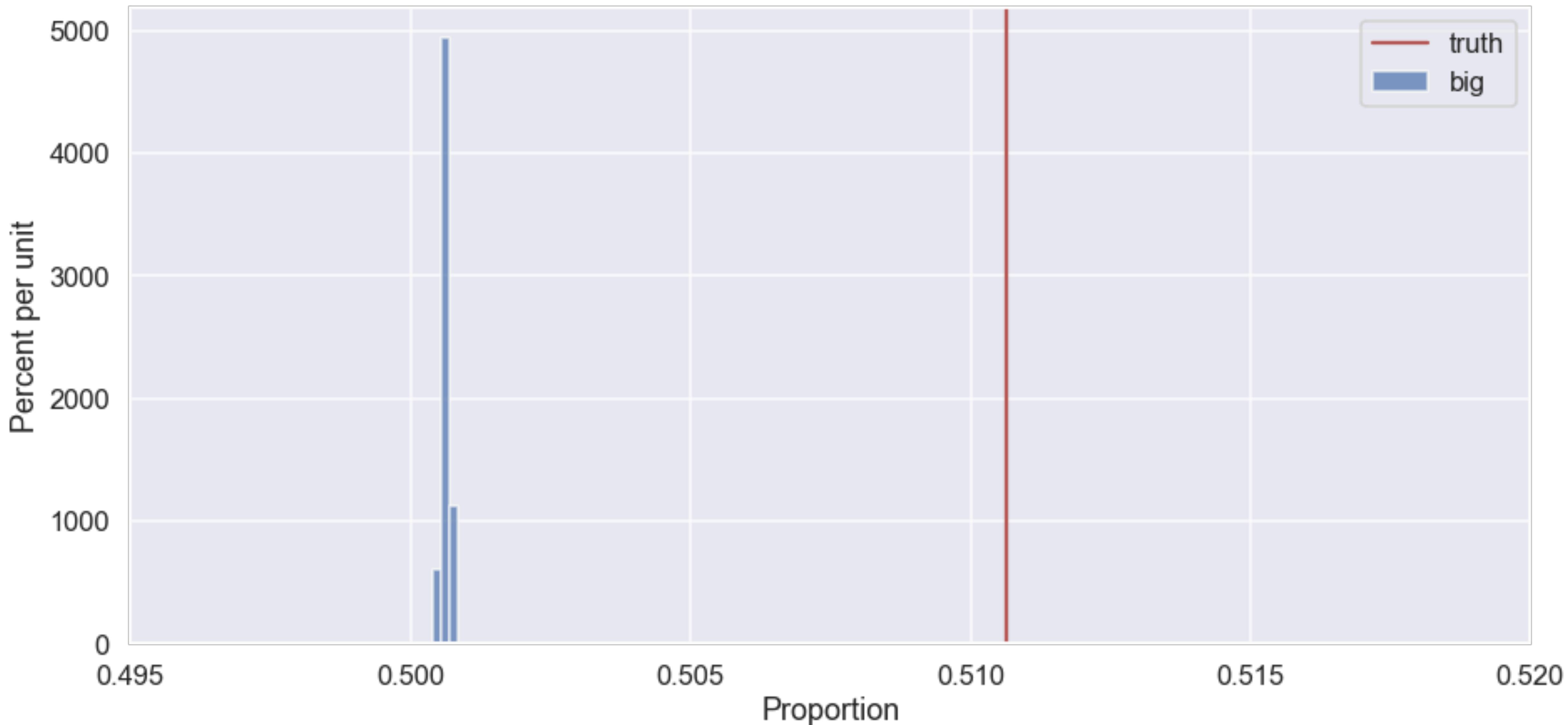
Non-random sample of 99% of U.S. voters

Create 10000 confidence intervals from a SRS of size 400.

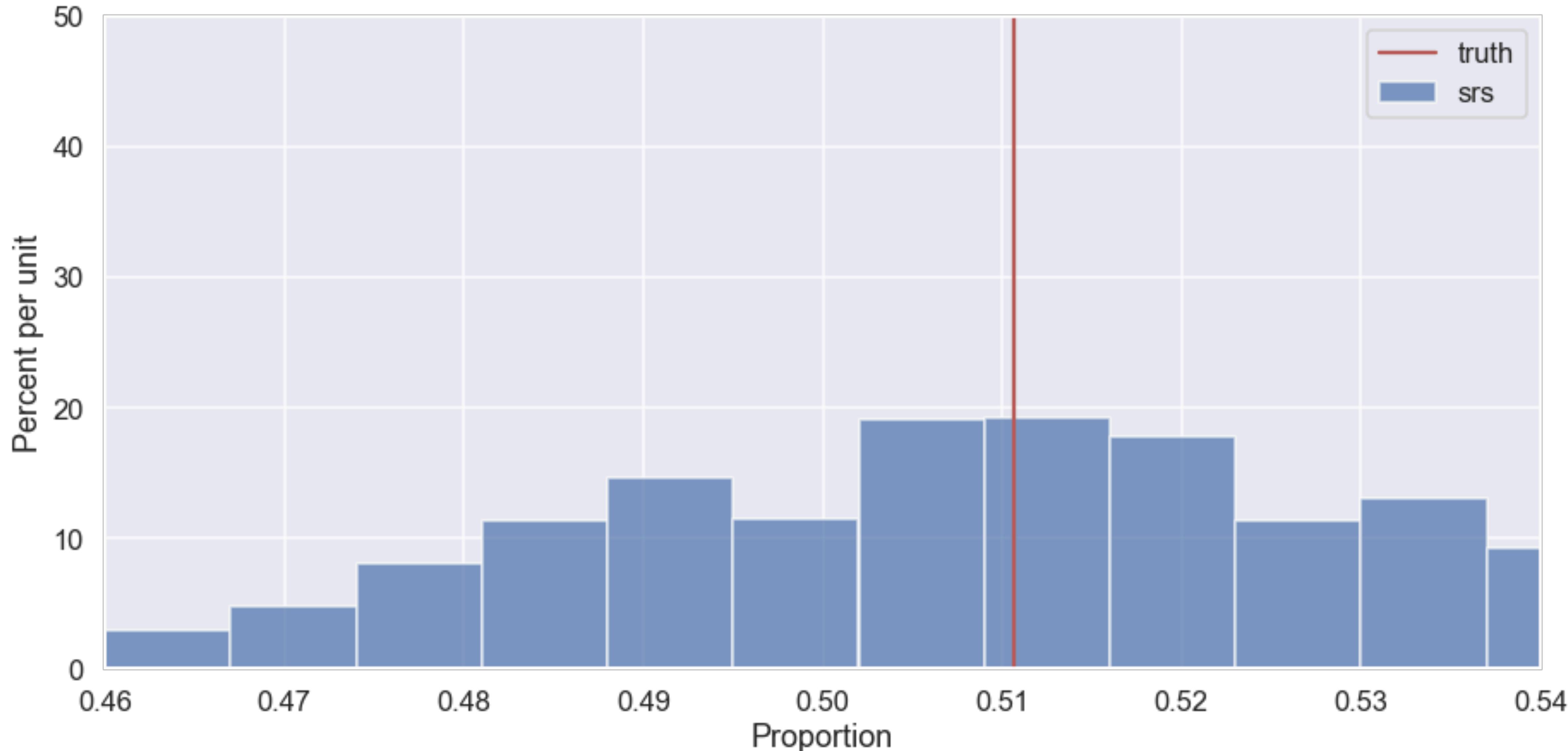
Create 10000 confidence intervals from a non-random sample of size 60mil with a 1% bias.

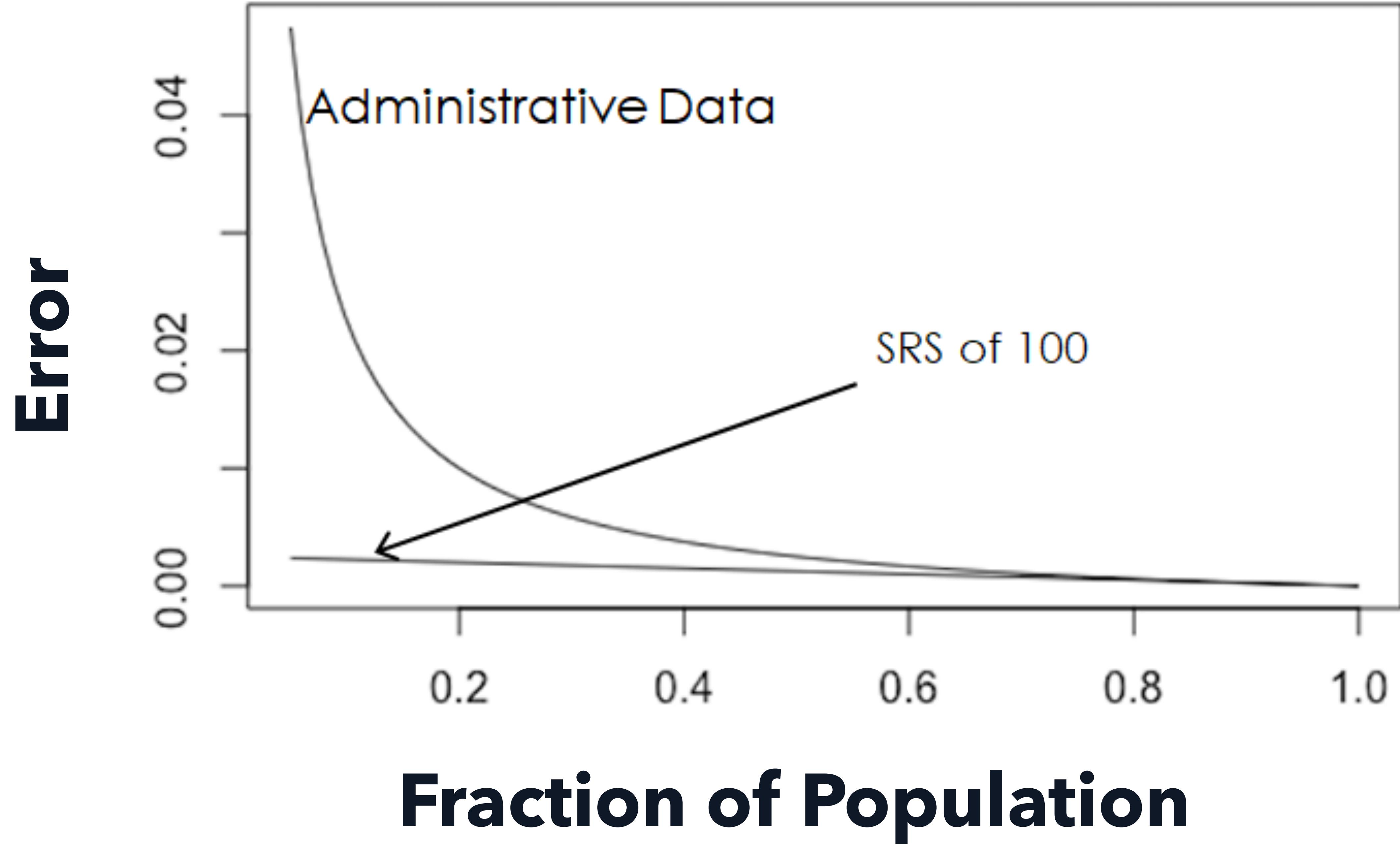
See which sample gives more accurate confidence intervals.

# "Big data"



# Simple Random Sample





# Summary

The **quality of a sample** matters a lot more than the quantity.

If a sample is non-random and biased, more data makes your predictions **worse!**

Use random samples or acknowledge potential sources of bias.

Don't estimate single numbers: **make confidence intervals.**