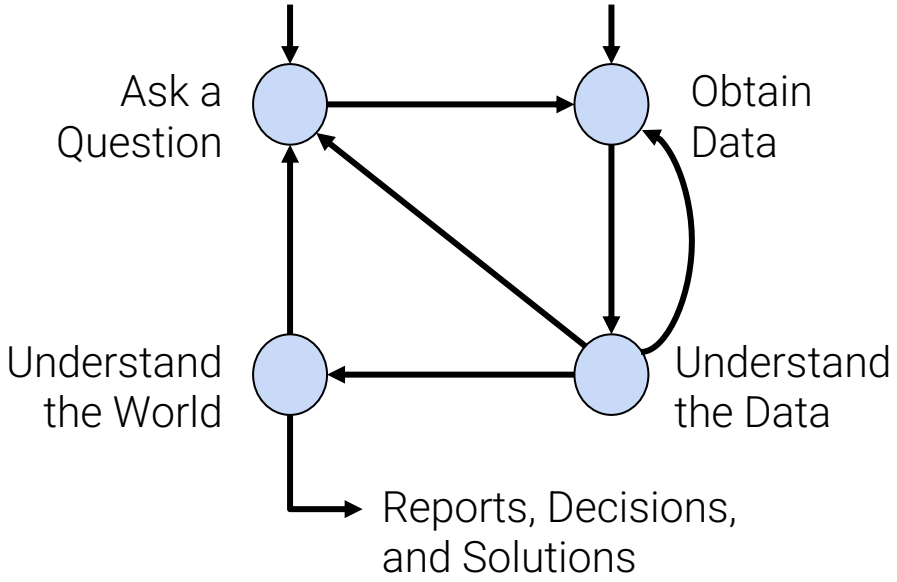
Sampling

How to sample effectively, and how to quantify the samples we collect.

Today's Roadmap

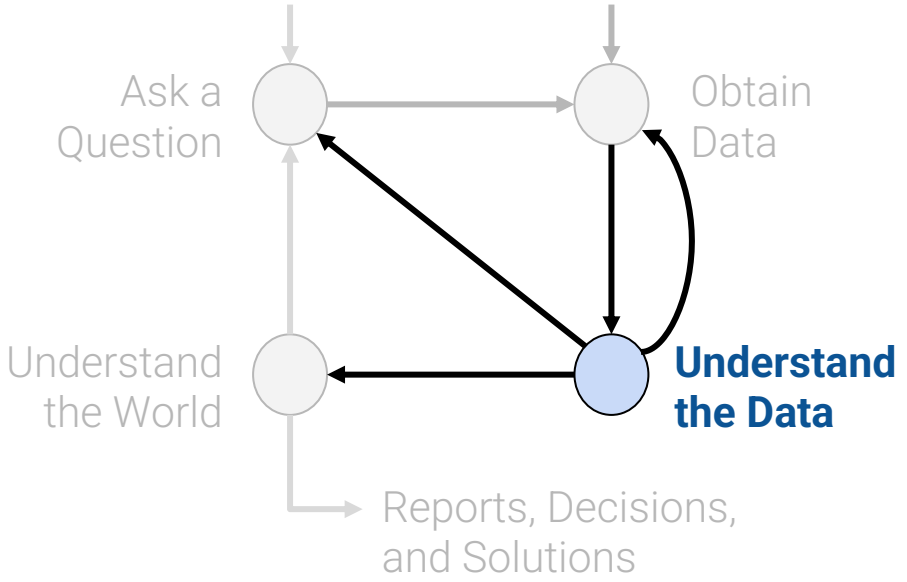
- Censuses and Surveys
- Sampling: Definitions
- Sampling Bias: A Case Study
- Probability Samples
- Multinomial Probabilities

Recall the
Data Science Lifecycle.



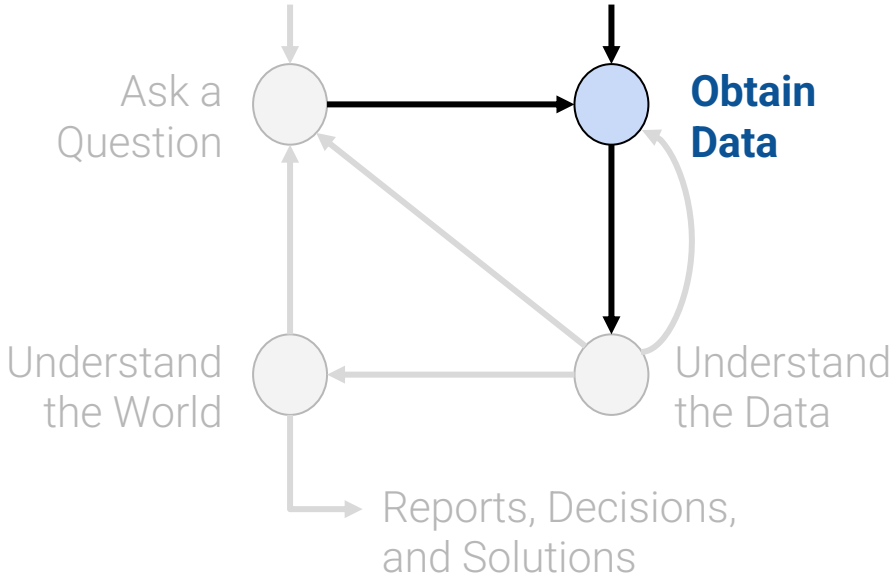
Before

We focused on Exploratory Data Analysis



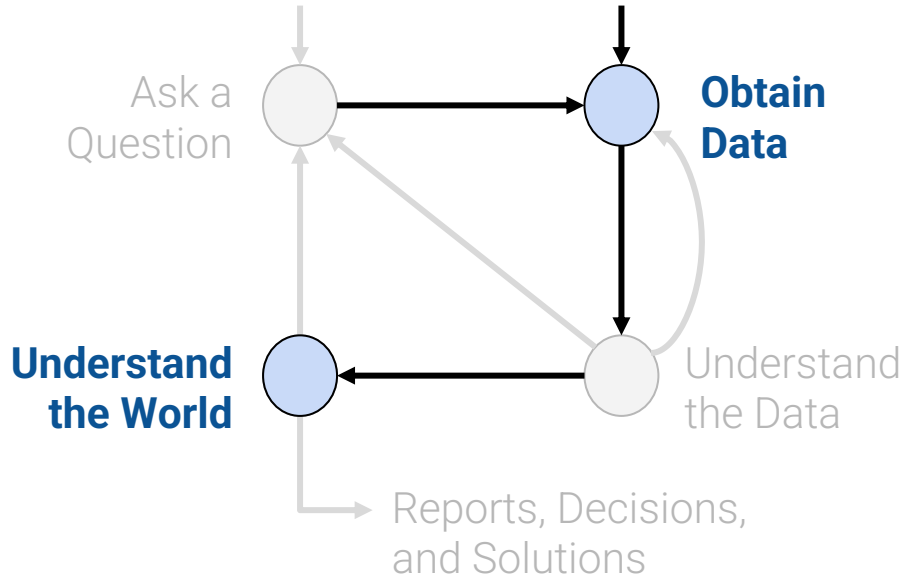
Today

How do we collect data?



Today

How does understanding data collection help us understand the world?



Censuses and Surveys

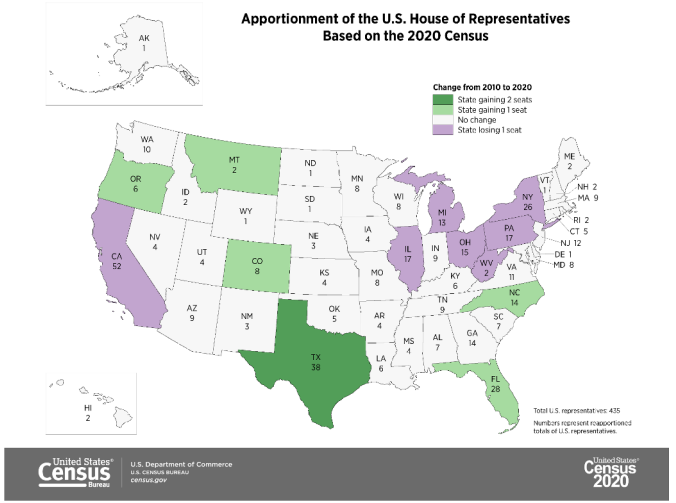
- **Censuses and Surveys**
- Sampling: Definitions
- Sampling Bias: A Case Study
- Probability Samples
- Multinomial Probabilities

Census

A **census** is “an official count or survey of a **population**, typically recording various details of individuals.”

The US Decennial Census

- Was last held in April 2020.
- Counts **every person** living in all 50 states, DC, and US territories. (Not just citizens.)
- Mandated by the Constitution. Participation is required by law.
- Important uses:
  - Allocation of Federal funds.
  - Congressional representation.
  - Drawing congressional and state legislative districts.



data.census.gov

Survey

Recall: a **census** is "an official count or **survey** of a population, typically recording various details of individuals."

A **survey** is a set of questions.

- For instance: census workers survey individuals and households.

What is asked, and how it is asked, can affect:

- How the respondent answers.
- **Whether** the respondent answers.

There are entire courses on surveying!

FiveThirtyEight

Politics Sports Science Podcasts Video Interactives

JUN. 27, 2019, AT 12:42 PM

The Supreme Court Stopped The Census Citizenship Question — For Now

NATIONAL

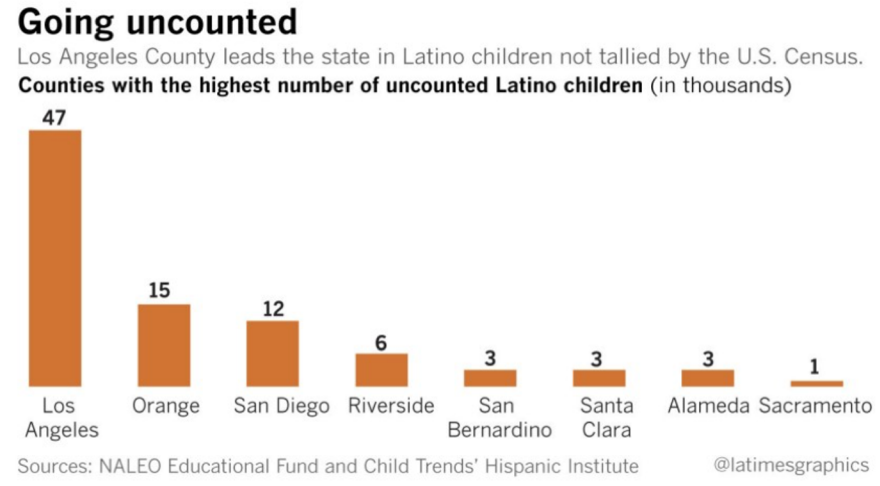
Citizenship Question To Be Removed From 2020 Census In U.S. Territories

August 9, 2019 · 3:23 PM ET

FiveThirtyEight, NPR

Undercounting in the US Decennial Census

LA Times 2010 Census



How do we know these numbers?
Other surveys

WaPo 2000 Census

High Court Rejects Sampling In Census
Ruling Has Political, Economic Impacts

Sampling methods would estimate Americans who missed the survey.
— Most often minorities/poor who vote Dem.
— "The better way is to improve the methods for contacting and questioning every household"

NY Times 2020 Census

In 2020 Census, Big Efforts in Some States. In Others, Not So Much.

California is spending $187 million to try to ensure an accurate count of its population. The Texas Legislature decided not to devote any money to the job. Why?

Sampling: Definitions

- Censuses and Surveys
- **Sampling: Definitions**
- Sampling Bias: A Case Study
- Probability Samples
- Multinomial Probabilities

Sampling from a finite population

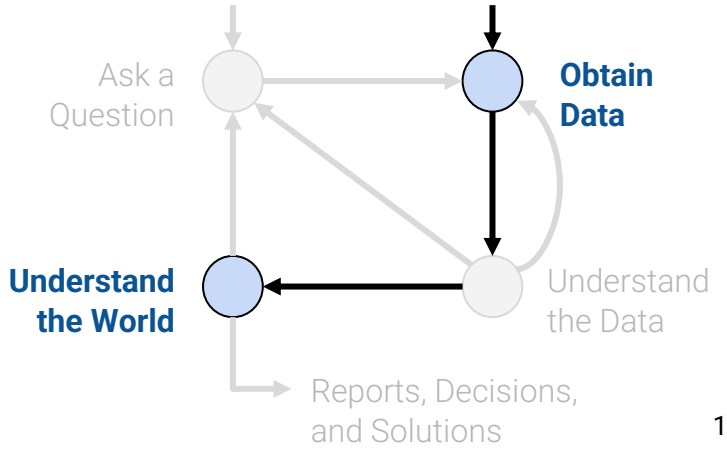
A census is great, but expensive and difficult to execute.

- Would **all** voters be willing to participate in a voting census prior to an actual election?

A **sample** is (usually) a subset of the population.

- Samples are often used to make **inferences about the population**.
- How you draw the sample will affect your accuracy.
- Two common sources of error:
  - **chance error**: random samples can vary from what is expected, in any direction.
  - **bias**: a systematic error in one direction.
    - Could come from our sampling scheme, and survey methods.

Inference: quantifying degree of certainty in our models of the world.



Population, sample, and sampling frame

Population: The group that you want to learn something about.

Sampling Frame: The list from which the sample is drawn.

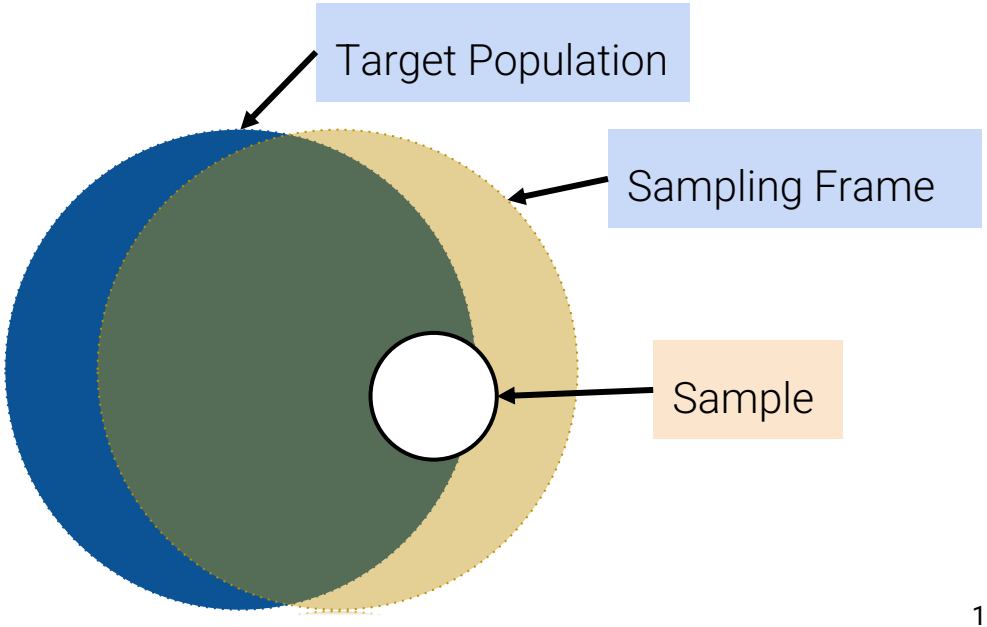
- If you're sampling people, the sampling frame is the set of all people that could possibly end up in your sample.

Sample: Who you actually end up sampling.

- A subset of your sampling frame.

There may be individuals in your **sampling frame** (and hence, your sample) that are **not** in your population!

Similarly, there might be individuals in your target population that are not in your sampling frame.



Other kinds of populations

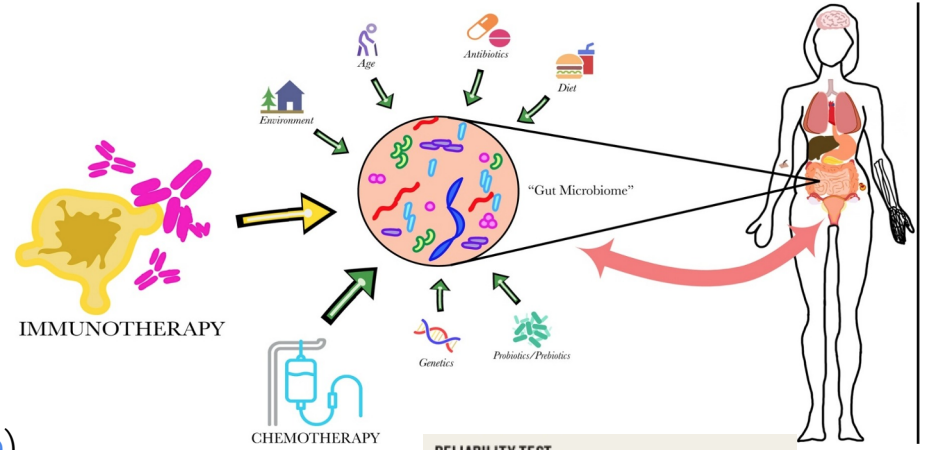
The individuals in a population are not always people!

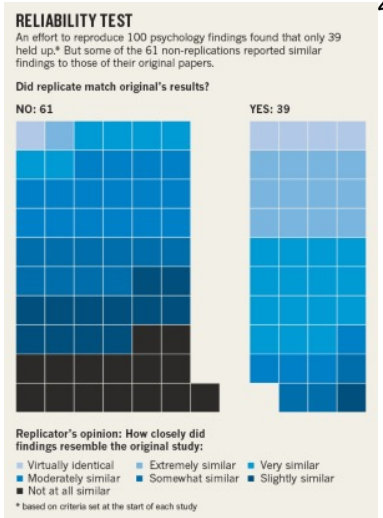
Could be

- **Bacteria** in your gut (sampled using DNA sequencing)
- **Trees** of a certain species
- **Small businesses** receiving a microloan
- **Published results** in a journal / field (example)

In any of these cases we might examine a sample and try to draw an inference about the population it came from.

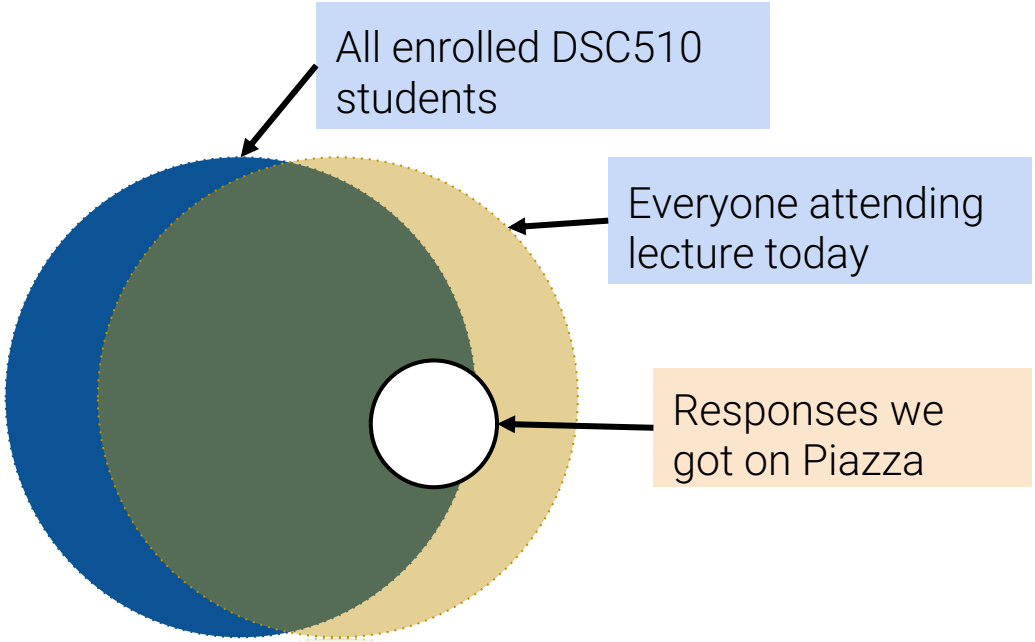
- Simplest example: what % have some binary property (like voting intention)?





In the last Slido question…

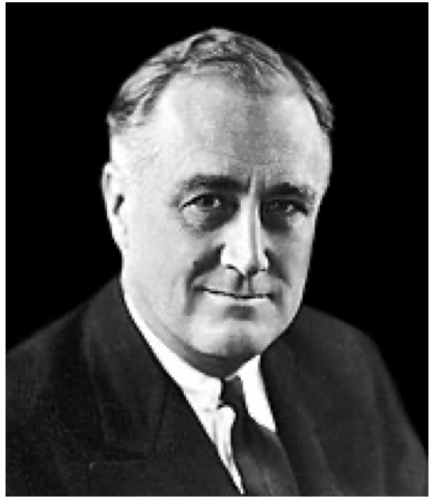
We want to study how DSC510 students engage with lectures.



Bias: A Case Study

- Censuses and Surveys
- Sampling: Definitions
- **Sampling Bias: A Case Study**
- Probability Samples
- Multinomial Probabilities

Case study: 1936 Presidential Election



Roosevelt (D) **Landon (R)**

In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, **polls** were conducted in the months leading up to the election to try and predict the outcome.

(Election result spoiler: Landon was not a U.S. President)

The Literary Digest: Election Prediction

The *Literary Digest* was a magazine. They had successfully predicted the outcome of 5 general elections coming into 1936.

They sent out their survey to **10,000,000** individuals, who they found from:

- Phone books.
- Lists of magazine subscribers.
- Lists of country club members.

	% Roosevelt	**# surveyed**
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

How could this have happened?
They surveyed 10 million people!

The Literary Digest: What happened?

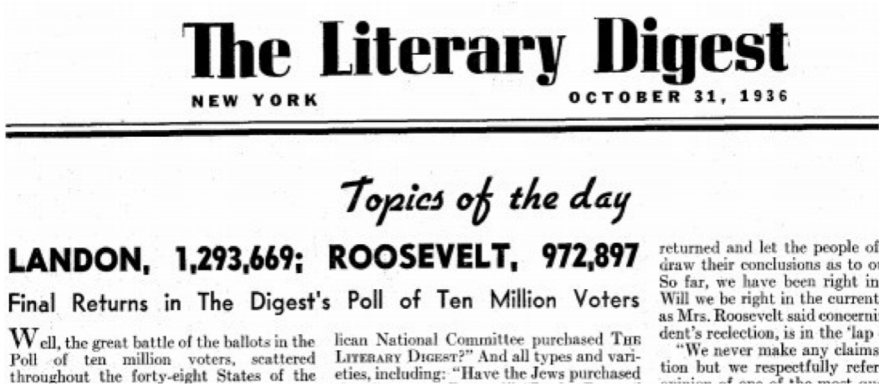
(1) The Literary Digest sample was **not representative** of the population.

- The Digest's **sampling frame**: people in the phonebook, subscribed to magazines, and went to country clubs.
- These people were more affluent and tended to vote Republican (Landon).

	% Roosevelt	# surveyed
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

(2) Only 2.4 million people **actually filled out the survey!**

- 24% response rate (low).
- Who knows how the 76% **non-respondents** would have polled?



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

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

returned and let the people of draw their conclusions as to o So far, we have been right in Will we be right in the current as Mrs. Roosevelt said concerni dent's reelection, is in the 'lap "We never make any claims tion but we respectfully refer

Gallup's Poll: Election Prediction

George Gallup, a rising statistician, also made predictions about the 1936 elections.

His estimate was **much** closer despite having a smaller **sample size** of "only" 50,000

(Also more than necessary!)

	% Roosevelt	# surveyed
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000
George Gallup's poll	56%	50,000
George Gallup's prediction of Digest's prediction	44%	3,000

George Gallup also predicted what The Literary Digest was going to predict, within 1%, with a **sample size of only 3000 people**.

- He predicted the Literary Digest's **sampling frame** (phonebook, magazine subscribers, country clubs).
- So he sampled those same individuals!

Samples, while convenient, are subject to chance error and **bias**.

Adapted from *Statistics* by Freedman, Pisani, and Purves, W.W. Norton NY

Common Biases

Selection Bias

- Systematically excluding (or favoring) particular groups.
- **Example**: The Literary Digest poll excludes people not in phone books.
- **How to avoid**: Examine the sampling frame and the method of sampling.

Response Bias

- People don't always respond truthfully.
- **Example**: Asking citizenship questions on the census survey→illegal immigrants might not answer truthfully
- **How to avoid**: Examine the nature of questions and the method of surveying.

Non-response Bias

- People don't always respond → People who don't respond aren't like the people who do!
- **Example**: Only 2.4m out of 10m people responded to The Literary Digest poll.
- **How to avoid**: Keep your surveys short, and be persistent.

The Gallup Poll today

Gallup U.S. Election polls:

- **Sampling Frame**: "civilian, non-institutionalized population" of adults in telephone households in continental US
- **Random Digit Dialing** to include both listed/unlisted phone numbers (to avoid **selection bias**)
- **Within household selection process** to randomly select if ≥1 adult in household
  - If no answer, recall multiple times (to avoid **non-response bias**)

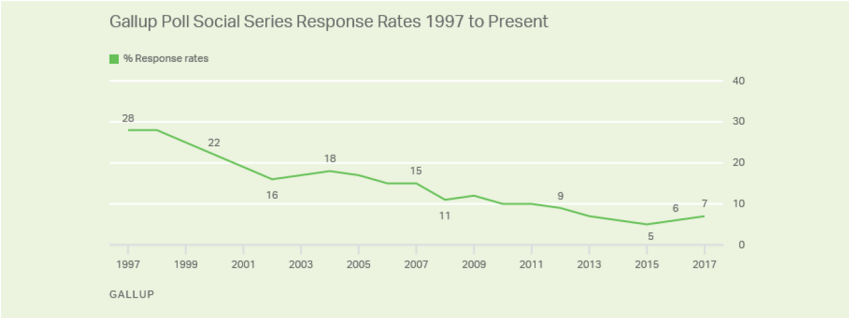
Gallup Poll

Election polling is hard!

Many sources of bias:

- **Who responds** to polls?
- Do voters **tell the truth?**
- How can we **predict turnout?**

Single-digit response rates are the norm

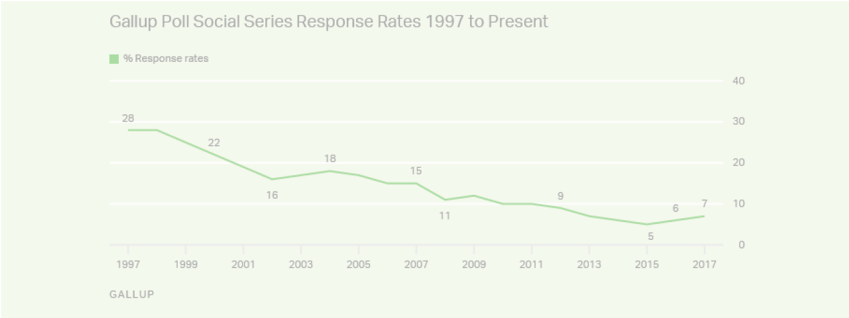


Election polling is hard!

Many sources of bias:

- **Who responds** to polls?
- Do voters **tell the truth?**
- How can we **predict turnout?**

Single-digit response rates are the norm



Poll numbers we see in the news are filtered through **proprietary statistical algorithms** that re-weight respondents
→ **"house effects"** of different pollsters

2022 ELECTION

Will The Polls Overestimate Democrats Again?

By Nate Silver
SEP. 16, 2022, AT 6:00 AM

FiveThirtyEight (source)

Polling bias isn't consistent

Weighted-average statistical bias in polls in final 21 days of the campaign

	PRES.	STATE LEVEL			
CYCLE	GENERAL	GOVERNOR	U.S. SENATE	U.S. HOUSE	COMBINED
1998	–	R+5.8	R+4.5	R+0.9	R+3.8
1999-2000	R+2.4	R+0.2	R+2.8	D+1.2	R+1.8
2001-02	–	D+3.5	D+2.0	D+1.4	D+2.6
2003-04	D+1.1	D+1.9	D+0.8	D+2.1	D+1.4
2005-06	–	D+0.4	R+2.1	D+1.1	D+0.1
2007-08	D+1.0	R+0.1	D+0.1	D+1.4	D+0.9
2009-10	–	R+0.2	R+0.8	D+1.3	D+0.4
2011-12	R+2.5	R+1.6	R+3.1	R+3.2	R+2.8
2013-14	–	D+2.3	D+2.7	D+3.9	D+2.8
2015-16	D+3.3	D+3.1	D+2.8	D+3.4	D+3.0
2017-18	–	R+0.9	EVEN	R+0.8	R+0.5
2019-20	D+4.2	D+5.6	D+5.0	D+6.1	D+4.8
All years	D+1.3	D+0.9	D+0.7	D+1.2	D+1.1

Probability Samples

- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- **Probability Samples**
- Multinomial Probabilities

Quality, not quantity!

A **huge sample size** does not fix a **bad sampling method**!

We want the sample to be **representative** of the population.

Think about **tasting soup**: if it's **well-stirred**, a spoonful is all you need!

- Don't just try to get a BIG sample. If your method of sampling is BAD, and your sample is BIG, what you'll have is a BIG BAD sample

Easiest way to to get a representative sample is by using **randomness**.

Non-random sample: Convenience Samples

(source)

An example of a non-random sample is **convenience sample.** It's whatever we can get ahold of.

Example: Scientists in New South Wales (AUS) collect specimens from eucalyptus trees to keep in museums, recording **where they came from** in latitude / longitude.

*Can we use this data to map the **geographic distribution** of eucalyptus trees?*

Convenience Samples

(source)

An example of a non-random sample is **convenience sample.** It's whatever we can get ahold of.

Example: Scientists in New South Wales (AUS) collect specimens from eucalyptus trees to keep in museums, recording **where they came from** in latitude / longitude.

*Can we use this data to map the **geographic distribution** of eucalyptus trees?*

Warning:

- Haphazard ≠ **random**.
- Many potential sources of bias!

Like polls, we can try to correct bias by statistical modeling. But better if we don't have to!

Probability Sample (aka Random Sample)

Why sample at random?

1. (As mentioned before) To get more representative samples → **reduce bias**
    - Random samples **can** produce biased estimates of population quantities.
2. More importantly, with random samples we can **estimate** the **bias** and **chance error** → **quantify uncertainty**

Probability Sample (aka Random Sample)

Why sample at random?

1. (As mentioned before) To get more representative samples → **reduce bias**
   - Random samples **can** produce biased estimates of population quantities.
2. More importantly, with random samples we can **estimate** the **bias** and **chance error** → **quantify uncertainty**

For a **probability sample**,

- We have to be able to provide the **chance** that any specified **set** of individuals will be in the sample.
- All individuals in the population **need not** have the same chance of being selected.
- Because we know all the probabilities, we will be able to **measure the errors**.

The real world is usually more complicated!

- Election polling: When Gallup calls, most people don't answer.
- Bacteria: We don't know the probability a given bacterium will get into a microbiome sample.

If the sampling / measurement process isn't fully under our control, we try to **model it**.

Example Scheme 1: Probability Sample

Suppose I have 3 TA's (**A**lan, **B**ennett, **C**eline):
I decide to sample 2 of them as follows:

- I choose **A** with probability 1.0
- I choose either **B** or **C**, each with probability 0.5.

All subsets of 2:	{**A**, **B**}	{**A**, **C**}	{**B**, **C**}
Probabilities:	0.5	0.5	0

This is a **probability sample** (though not a great one).

- Of the 3 people in the population, I know the chance of getting each subset.
- Suppose I'm measuring the average distance TA's live from campus.
  - This scheme does not see the entire population!
  - My estimate using the single sample I take has some **chance error** depending on if I see AB or AC.
  - This scheme **biases** towards A's response

Common random sampling schemes

A **random sample with replacement** is a sample drawn **uniformly** at random **with** replacement.

- Random doesn't always mean "uniformly at random," but in this specific context, it does.
- Some individuals in the population might get picked more than once

A **simple random sample (SRS)** is a sample drawn **uniformly** at random **without** replacement.

- **Every individual (and subset of individuals) has the same chance of being selected.**
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

A raffle could use either sampling scheme, depending on if winners are eligible for multiple prizes.

Example Scheme 2: Simple Random Sample?

We have the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

2. Does each student have the same probability of being selected?

3. Is this a simple random sample?

Example Scheme 2: Simple Random Sample?

Consider the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random (e.g. Student 8).
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

Yes.

For a sample [n, n + 10, n + 20, …, n + 1090], where 1 <= n <= 10, the probability of that sample is 1/10.

Otherwise, the probability is 0.

Only 10 possible samples!

2. Does each student have the same probability of being selected?

Yes.

Each student is chosen with probability 1/10.

3. Is this a simple random sample?

No.

The chance of selecting (8, 18) is 1/10; the chance of selecting (8, 9) is 0.

Barbie vs. Oppenheimer

On July 21st, two highly anticipated movies arrived in theaters: **Barbie** and **Oppenheimer**.

We want to know which movie will prevail on opening day, in Berkeley.

NY Times, GQ

Demo

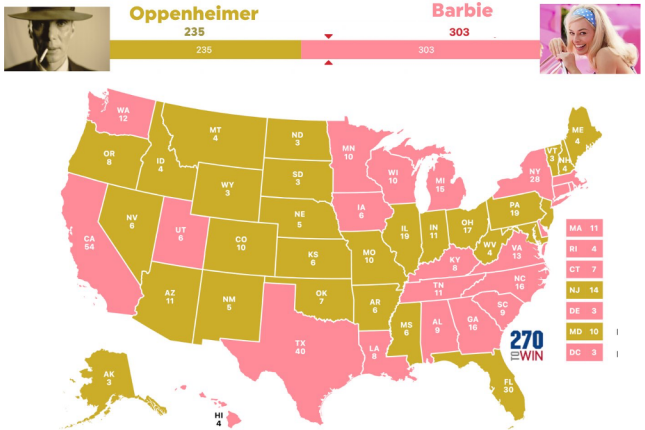
Imaginary Barbie Land Pollster

Suppose we took a sample of Berkeley residents to predict the box office outcomes.

- We poll all **retirees** for their preference.
- Even if they answer truthfully, this is a **convenience sample**.

Then, suppose July 21st happens.

- How "off" is our sample estimate from the actual outcome?
- How would a random sample with replacement have performed?



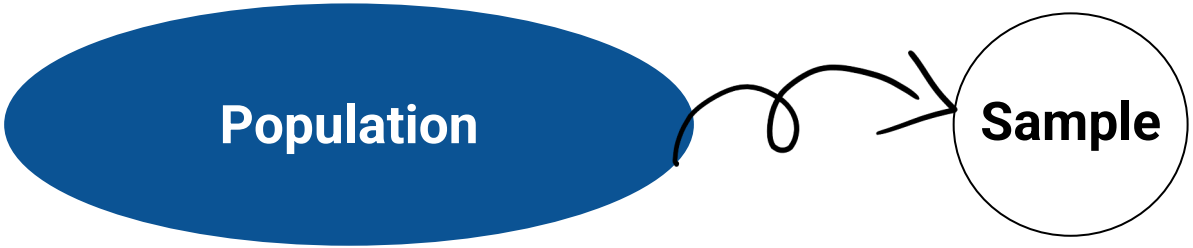
Twitter

Demo

Multinomial Probabilities

- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- **Multinomial Probabilities**

So far

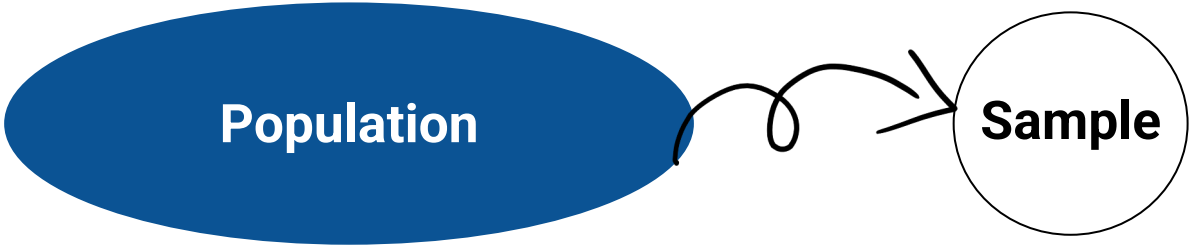


If we have a probability sample:

- We can quantify error and bias.
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

Note: We almost **never** know the population distribution! But this is a good start.

So far



If we have a probability sample:

- We can quantify error and bias (to be covered later).
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

Note: We almost **never** know the population distribution! But this is a good start.

Special case: Random sampling with replacement of a **categorical population** produces **Multinomial Probabilities**.

A very common approximation for sampling

A common situation in data science:

- We have an enormous population.
- We can only afford to sample a relatively small number of individuals.

If the **population is huge** compared to the sample, then
random sampling **with** and **without** replacement are pretty much the **same**.

Example: Suppose there are 10,000 people in a population.
Exactly 7,500 of them like Reese's; the other 2,500 like Snickers.

What is the probability that in a random sample of 20, **all people like Reese's**?

SRS (Random Sample Without Replacement)

$$\overbrace{\frac{7500}{10000}}^{0.75}\overbrace{\frac{7499}{9999}}^{0.74997}\cdots\overbrace{\frac{7482}{9982}}^{0.7495}\overbrace{\frac{7481}{9981}}^{0.7495} \approx .003151$$

Random Sample With Replacement

$$\left(\frac{7500}{10000}\right)^{20} \approx .003171$$

Probabilities of sampling with replacement are much easier to compute!

Drawing samples from categorical distribution

Multinomial probabilities arise when we sample:

- Uniformly at random, **with replacement**.
- A fixed number (n) times.
- From a **categorical distribution**.
  - If 2 categories, **Binomial**
  - If > 2 categories, **Multinomial**:

    Bag of Marbles: 60% **blue** marbles 30% are **green** 10% are **red**

Goal: **Count the number of each category** that end up in our sample.

- If we draw 7 marbles from this bag of marbles, how many of each color will we get?
- We can simulate using NumPy
  - `np.random.multinomial` returns these counts.

Multinomial Probabilities

Suppose we get a sample of 4 **blue** marbles, 2 **green** marbles, and 1 **red** marble. What is the probability of getting this sample?

Q1. What is P(**bgbbbgr**)?

Use product rule to determine probability for a particular **order**:

$$P(\textbf{bgbbbgr}) = 0.6 \times 0.3 \times 0.6 \times 0.6 \times 0.6 \times 0.3 \times 0.1 = (0.6)^4(0.3)^2(0.1)^1$$

=

Q2. What is P(4 **blue**, 2 **green**, 1 **red**)?

$$\frac{7!}{4!\,2!\,1!}(0.6)^4(0.3)^2(0.1)^1$$

multinomial probability

We use the **addition rule** and **multiplication rule**:

# of ways to choose 4 of 7 places to write **b**, then choose 2 places to write **g** , (other 1 get filled with **r**)

For a particular outcome (say, Q1), probability of this **ordered series** of **b**'s, **g** 's, and **r**'s

Multinomial Probabilities: generalized

If we are drawing at random with replacement **n** times, from a population broken into three separate categories (where $p_1 + p_2 + p_3 = 1$):

- Category 1, with proportion $\mathbf{p_1}$ of the individuals.
- Category 2, with proportion $\mathbf{p_2}$ of the individuals.
- Category 3, with proportion $\mathbf{p_3}$ of the individuals.

Then, the **multinomial probability** of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

$$\frac{n!}{k_1!k_2!k_3!}p_1^{k_1}p_2^{k_2}p_3^{k_3}$$

At no point in this class will you be required to memorize this! This is just for your own understanding.

Summary

Understanding the sampling process is what lets us go from **describing the data** to **understanding the world**

Without knowing / assuming something about how the data were collected:

- There is no connection between the **sample** and the **population**
- The **data set** doesn't tell us about the **world behind the data**