slido

1760929

Join at slido.com
#1760929

ⓘ Click **Present with Slido** or install our Chrome extension to display joining instructions for participants while presenting.

CC BY NC SA

⚠️ Reminder to start the Zoom recording!

1760929

📞 Phone chats w/ Josh: Lots of slots in March

LECTURE 9

Sampling

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

Data 100/Data 200, Spring 2025 @ UC Berkeley

Narges Norouzi and Josh Grossman

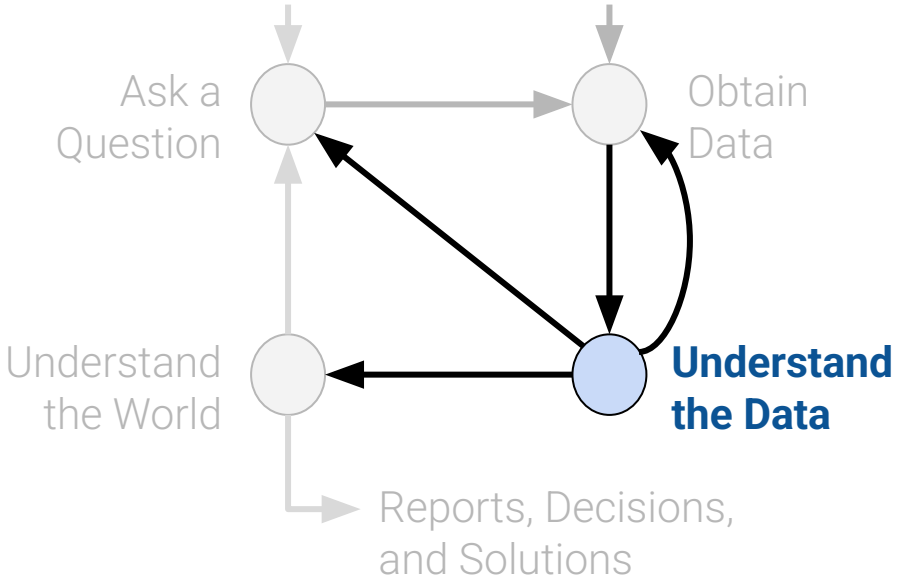
Content credit: Acknowledgments

CC BY NC SA

1760929

Before

We focused on EDA

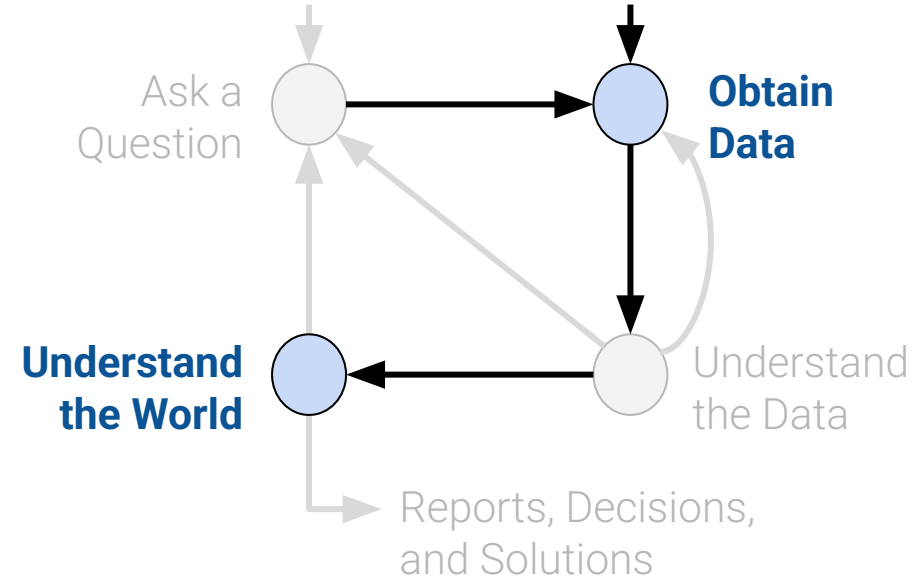


1760929

Today

How do we collect data?

How does understanding data collection help us understand the world?



1760929

Today's Roadmap

Lecture 9, Data 100 Spring 2025

- Censuses and Surveys
- Sampling: A Case Study
- Sampling Errors
- Types of Sampling
- Post-stratification

CC BY NC SA

1760929

Censuses and Surveys

Lecture 9, Data 100 Spring 2025

- **Censuses and Surveys**
- Sampling: A Case Study
- Sampling Errors
- Types of Sampling
- Post-stratification

CC BY NC SA

A Census

1760929

A **census** is a **complete** count or survey of a **population**.

- **Every individual** is included!

The **population** is the complete set of studied individuals.

Example populations:

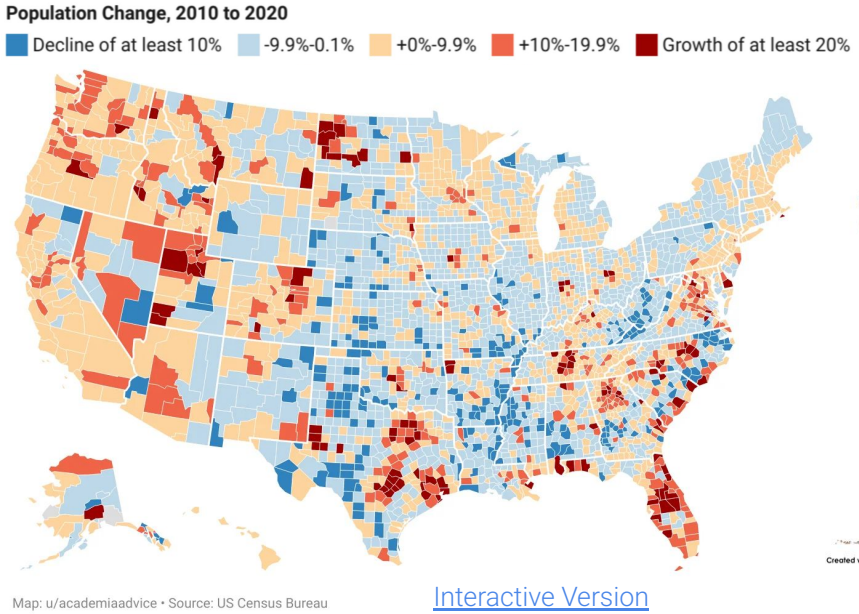
- **People** living in a particular country
- **Bacteria** in a person's gut
- **Trees** of a certain species

CC BY NC SA

1760929

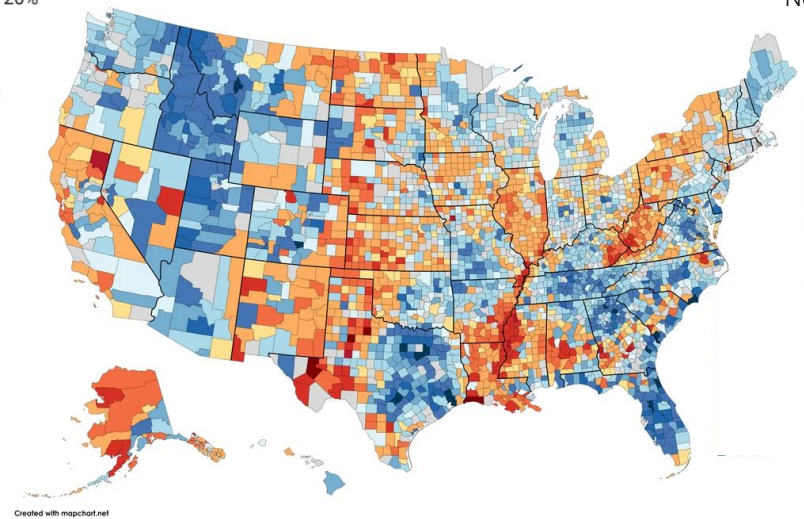
The US Decennial Census

- Last held in 2020; next one in 2030
- *Attempts* to count **every person** living in all 50 states, DC, and US territories.



Interactive Version

Between decades: **American Community Survey (ACS)**. A sample, not a census!



Source

CC BY NC SA

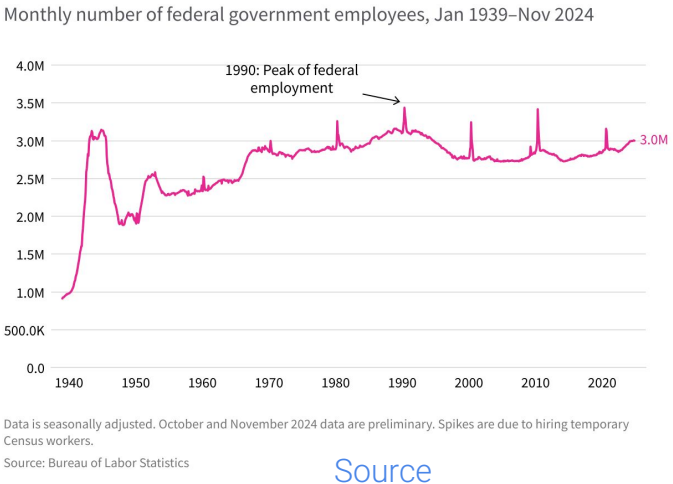
Survey

1760929

Recall: A **census** is "a complete count or **survey** of a population."

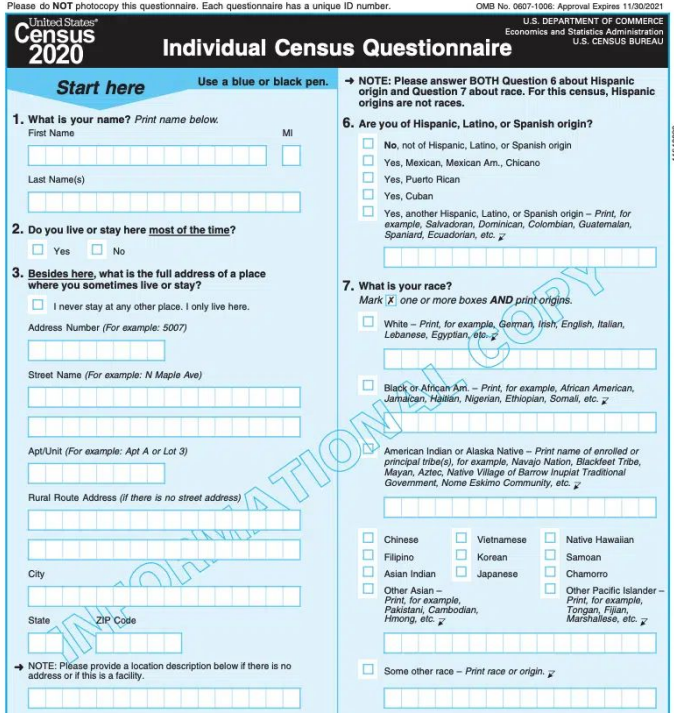
- A **survey** is a set of questions or measurements.

Stat 152 @ UC Berkeley (Sampling Surveys)



Source

A 2020 Census Taker (source)



2020 Census Form (source)

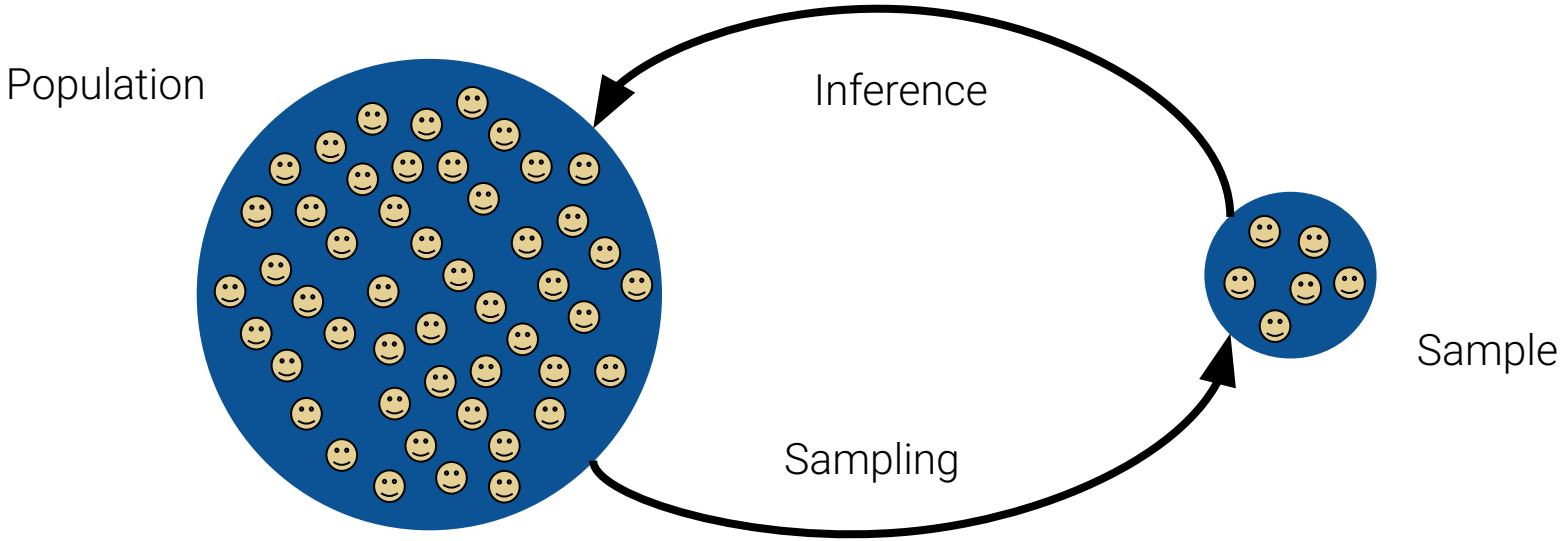
Sampling

1760929

A census is ideal, but **expensive** and **difficult** to execute.

A **sample** is a subset of a population.

- **Inference**: Drawing conclusions about a population based on a sample.



Note: We often think of a population as having infinite size, or as a data-generating **process**. More on this when we get to inference!

1760929

Sampling: A Case Study

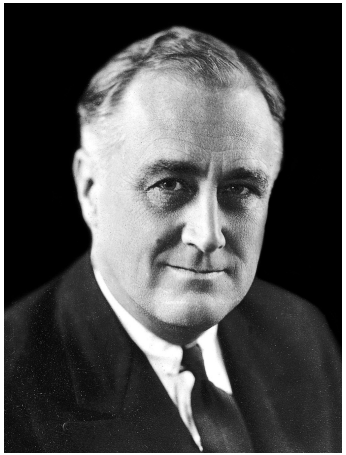
Lecture 9, Data 100 Spring 2025

- Censuses and Surveys
- **Sampling: A Case Study**
- Sampling Errors
- Types of Sampling
- Post-stratification

CC BY NC SA

Case study: 1936 Presidential Election

1760929



Roosevelt (Democrat)

Landon (Republican)

In 1936, President **Franklin D. Roosevelt** went up for re-election against **Alf Landon**.

Election polls were conducted to try and predict the outcome.

CC BY NC SA

The Literary Digest: 1936 Election Prediction

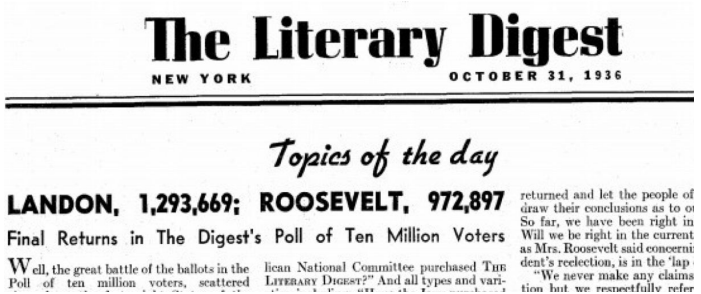
1760929

A magazine called ***Literary Digest*** successfully predicted the outcome of 5 presidential elections before 1936.

They sent a survey to **10,000,000 (!)** individuals, using contact info from:

- Phone books.
- Literary Digest subscribers.
- Automobile registrations.

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



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

lican National Committee purchased THE LITERARY DIGEST?" And all types and vari-

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

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

CC BY NC SA

The Literary Digest: What happened?

1760929

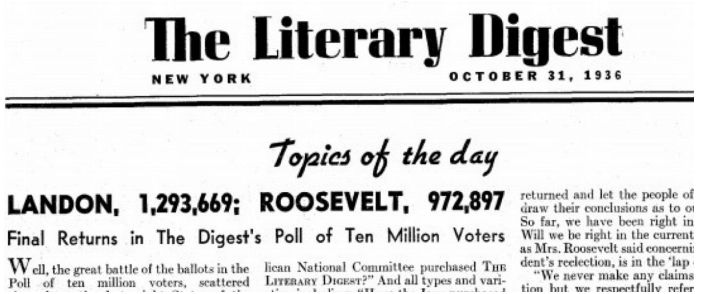
(1) The Literary Digest sample was **not representative** of the population.

- The sampled voters were **more affluent** and **tended to vote Republican** (Landon).

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

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

- 24% response rate
- 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

lican National Committee purchased THE LITERARY DIGEST?" And all types and vari-

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

CC BY NC SA

George Gallup's Poll: Election Prediction

1760929

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

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

Gallup's secret sauce: A more **representative random sample**.

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

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

1760929

The best way to get a representative sample is through ***randomization***.

(Though, sometimes it's not easy or possible.)

CC BY NC SA

1760929

Sampling Errors

Lecture 9, Data 100 Spring 2025

- Censuses and Surveys
- Sampling: A Case Study
- **Sampling Errors**
- Types of Sampling
- Post-stratification

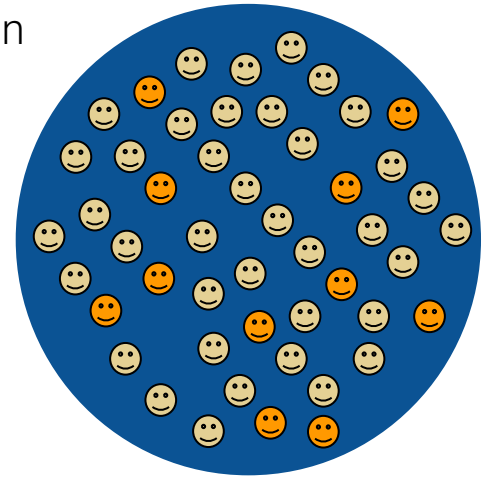
CC BY NC SA

1760929

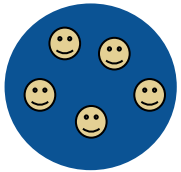
Chance error (i.e., variance)

- Random samples can vary from what is expected, in any direction.
- One way to reduce: Increase size of random sample.
- Another option: Stratify. More on this soon!

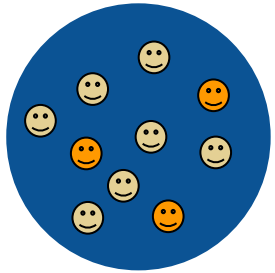
Population



Random sample n=5



Random sample n=10



Larger random sample is more likely to be **representative.**
[i.e., less likely to get an "unlucky" draw]

CC BY NC SA

Sources of Error in Sampling

1760929

Chance error (i.e., variance)

- Random samples can vary from what is expected, in any direction.
- One way to reduce: Increase size of random sample.
- Another option: Stratify. More on this soon!

Bias

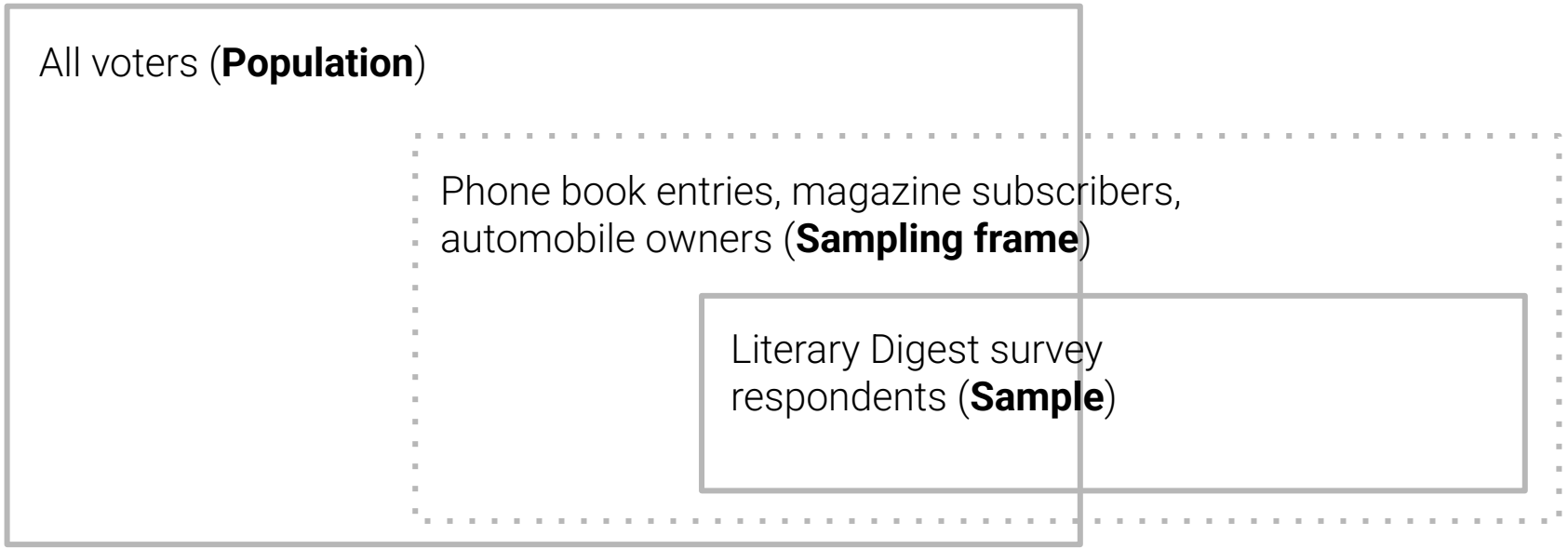
- A systematic error in one direction.
- Solution: Lots of possible sources, each with different reduction strategies.

CC BY NC SA

1760929

Selection Bias (i.e., sampling bias)

- Systematically excluding (or favoring) particular groups.
- **Example**: The Literary Digest poll excluded people not in phone books.

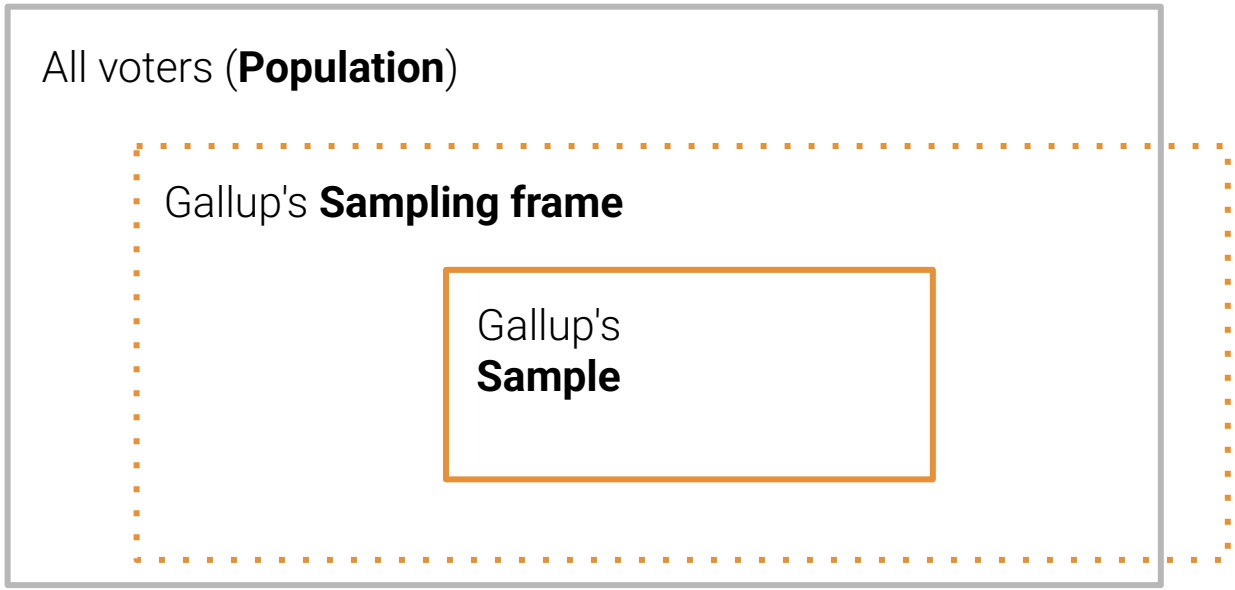


CC BY NC SA

1760929

Selection Bias (i.e., sampling bias)

- Systematically excluding (or favoring) particular groups.
- **Example**: The Literary Digest poll excluded people not in phone books.
- **How to avoid**: Randomly sample, and improve overlap of **sampling frame** and population.



CC BY NC SA

Common Biases

1760929

Response Bias (i.e., measurement bias)

- Miscalibrated survey questions. Desired measure differs from actual measure.
- **Obvious example**: "Will you vote for Roosevelt or Landon? If you say 'Roosevelt', I will give you $1."
- **Subtle example**: "Do you agree that you will vote for Roosevelt?" We tend to prefer agreeing over disagreeing.
- **How to avoid**: Improve questions. Lots of response bias subtypes+prevention methods.

Non-response Bias

- Survey respondents differ from non-respondents.
- **Example**: ~24% response rate to The Literary Digest poll.
- **How to avoid**: Increase response rate. For example, reduce the number and length of questions, incentivize completion, and follow up.

CC BY NC SA

Convenience Samples

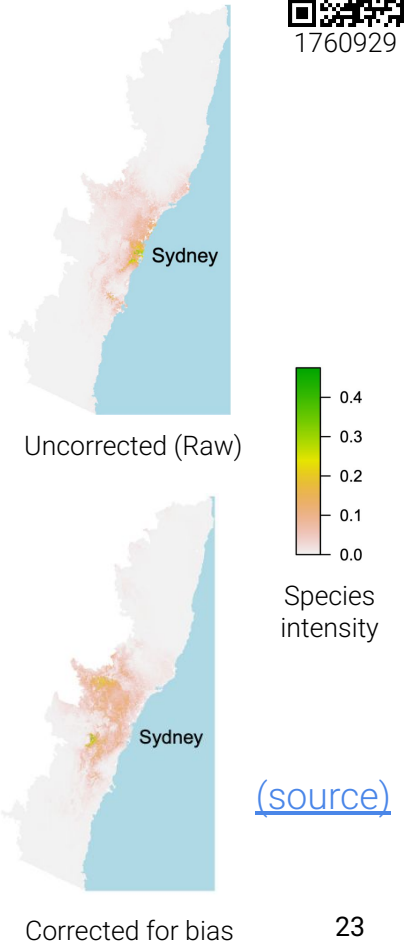
Convenience sample. Individuals we can easily access. Non-random!

Example: Scientists in New South Wales (Australia) 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: Selection bias!

People are also bad at mimicking true randomness.



Types of sampling

Lecture 9, Data 100 Spring 2025

- Censuses and Surveys
- Sampling: A Case Study
- Sampling Errors
- **Types of Sampling**
- Post-stratification

CC BY NC SA

Probability Samples

1760929

If we know the **probability** that any **subset** of individuals in the sampling frame will be selected, our sample is a **probability sample (i.e., a random sample).**

For example, suppose I have 3 TA's (**A**lan, **B**en, **C**eline). I want to sample **2 of them**.

Alan is always selected. P(A) = 100%


We flip a coin to pick Ben or Celine. P(B) = P(C) = 50%

P(A and B) = 50%

P(A and C) = 50%

P(B and C) = 0%

CC BY NC SA

Simple Random Sample (SRS)

1760929

Sample is drawn **uniformly** at random **without** replacement.

- Every subset of **n** individuals has the **same** chance of being selected, where n=1,2,3…

In other words:

- Every individual has the same chance as every other individual
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

SRS → "Names in a hat" sample

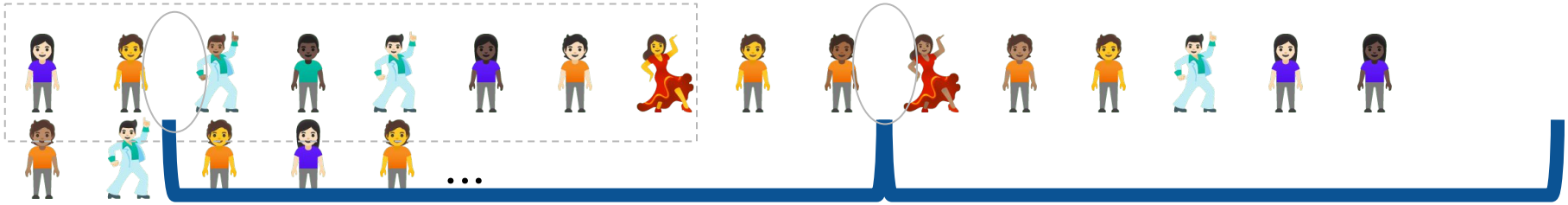
CC BY NC SA

Example Sampling Scheme: Simple Random Sample?

1760929

We have the following sampling scheme:

- A class roster has 1200 students listed alphabetically.
- We pick one of the first 10 students on the list at random (e.g., Student 3).
- To create the sample, pick the chosen student and every 10th student listed after that (e.g., Students 3, 13, 23, 33, …).



1. Is this a probability sample?

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

3. Is this a simple random sample (SRS)?

CC BY NC SA

1760929

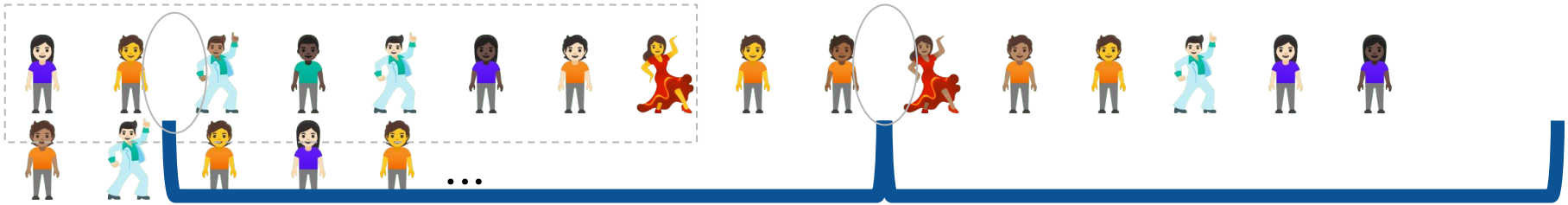
Is this a probability sample?

CC BY NC SA

slido

Example Sampling Scheme: Simple Random Sample?

1760929



1. Is this a probability sample?

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

3. Is this a simple random sample (SRS)?

Yes.

There are 10 possible samples. Each one is equally likely. All other combinations have probability 0.

CC BY NC SA

slido

1760929

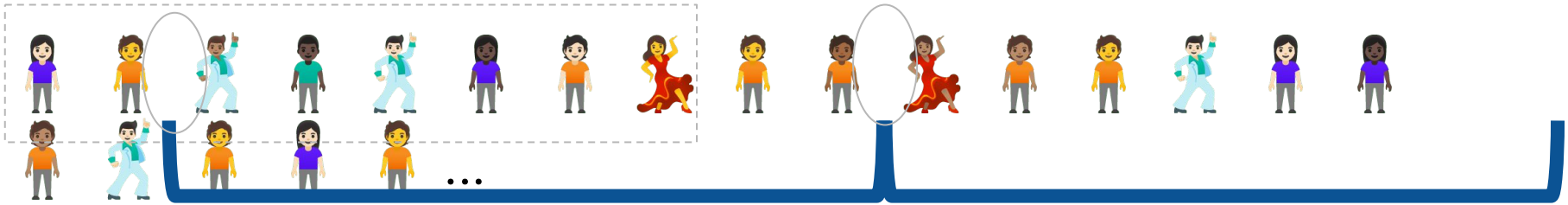
Does each student have the same probability of being selected?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

CC BY NC SA

Example Sampling Scheme: Simple Random Sample?

1760929



1. Is this a probability sample?

Yes.

There are 10 possible samples. Each one is equally likely. All other combinations have probability 0.

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 (SRS)?

CC BY NC SA

slido

1760929

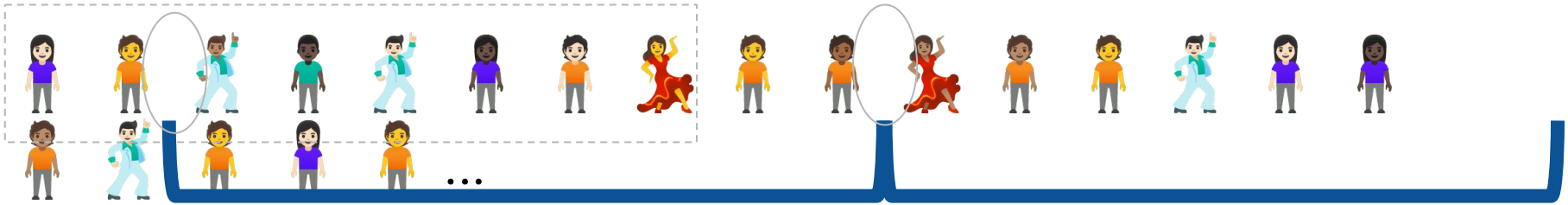
Is this a simple random sample?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

CC BY NC SA

Example Sampling Scheme: Simple Random Sample?

1760929



1. Is this a probability sample?

Yes.

There are 10 possible samples. Each one is equally likely. All other combinations have probability 0.

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 (SRS)?

No.

The chance of selecting (3, 13) is 1/10; the chance of selecting (3, 4) is 0.

CC BY NC SA

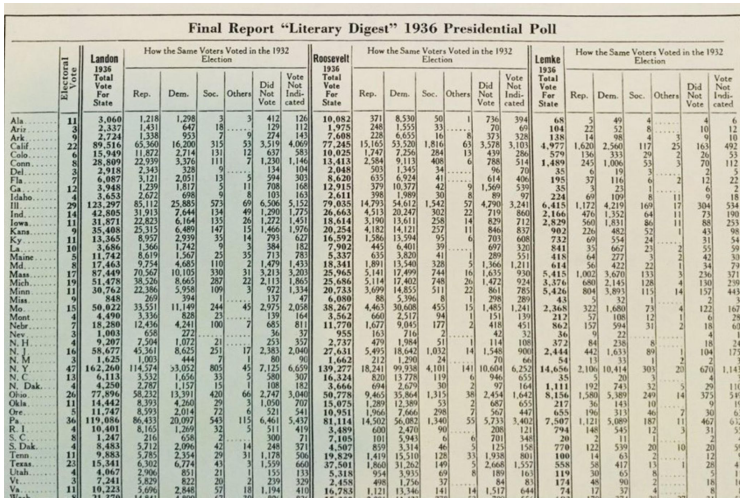
Simple Random Sample (SRS)

1760929

The 1936 Literary Digest poll predicted Roosevelt would lose the presidential election with 43% of the vote.

- In reality, Roosevelt won with 61% of the vote.

With a small representative sample of 1936 voters, could the election have been predicted accurately?



Final Report "Literary Digest" 1936 Presidential Poll

	Electoral Votes	Landon 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Roosevelt 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Lemke 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated
Ala.	11	3,060	1,218	1,298	3	3	412	126	10,082	371	8,530	50	1	736	394	68	5	49	4	....	4	6
Ariz.	3	2,337	1,431	647	18	....	129	112	1,975	248	1,555	33	....	70	69	104	22	52	8	....	10	12
Ark.	9	2,724	1,338	953	7	9	274	143	7,608	228	6,655	16	8	373	328	138	14	98	4	5	9	10
Calif.	22	89,516	65,360	16,200	315	53	3,519	4,069	77,245	15,165	53,520	1,816	63	5,578	3,103	4,977	1,620	2,560	117	25	163	492
Colo.	6	15,949	11,872	2,714	131	12	637	583	10,025	1,747	7,256	284	13	439	286	579	136	333	29	2	26	53
Conn.	8	28,809	22,939	3,376	111	7	1,230	1,146	13,413	2,584	9,113	408	6	788	514	1,489	245	1,006	53	3	70	112
Del.	3	2,918	2,343	328	9	....	134	104	2,048	503	1,345	34	....	96	70	35	6	19	3	....	2	5
Fla.	7	6,087	3,121	2,051	13	5	594	303	8,620	635	6,924	41	....	614	406	195	37	116	6	2	12	22
Ga.	12	3,948	1,239	1,817	5	11	708	168	12,915	329	10,372	42	9	1,569	539	35	3	23	1	....	6	2
Idaho	4	3,653	2,672	698	9	8	103	163	2,611	398	1,989	30	8	89	97	224	69	109	8	11	9	18
Ill.	29	123,297	85,112	25,885	573	69	6,506	5,152	79,035	14,793	54,612	1,542	57	4,790	3,241	6,415	1,172	4,219	169	17	304	534
Ind.	14	42,505	31,913	7,644	154	49	1,200	1,775	26,663	4,513	20,247	302	22	719	860	2,166	476	1,352	64	11	73	190
Iowa.	11	31,871	22,823	6,164	135	26	1,272	1,451	18,614	3,190	13,611	258	14	829	712	2,829	560	1,831	86	11	88	253
Kans.	9	35,408	25,315	6,409	147	15	1,466	1,976	20,254	4,182	14,121	257	11	846	837	902	226	482	52	1	43	98
Ky.	11	13,365	8,957	2,939	35	14	793	627	16,592	1,586	13,594	95	6	703	608	732	69	554	24	....	31	54
La.	10	3,686	1,366	1,742	9	3	384	182	7,902	445	6,401	39	....	697	320	841	35	662	23	2	55	59
Maine	5	11,742	8,619	1,567	25	35	713	783	5,337	635	3,820	41	1	289	551	418	64	272	3	2	42	30
Md.	8	17,463	9,754	4,685	110	2	1,479	1,433	18,341	1,891	13,540	328	5	1,366	1,211	614	56	422	22	1	34	79
Mass.	17	87,449	70,567	10,105	330	31	3,213	3,203	25,965	5,141	17,499	744	16	1,635	930	5,415	1,002	3,670	133	3	236	371
Mich.	19	51,478	38,526	8,665	287	22	2,113	1,865	25,686	5,114	17,402	748	26	1,472	924	3,376	680	2,145	128	4	130	299
Minn.	11	30,762	22,386	5,958	109	3	972	1,334	20,733	3,699	14,855	511	22	861	785	5,426	804	3,893	115	14	157	443
Miss.	9	848	269	394	1	....	157	47	6,000	88	5,396	8	1	298	209	43	5	32	1	....	4	1
Mo.	15	50,022	33,551	11,149	244	45	2,975	2,058	38,267	4,463	30,608	455	15	1,485	1,241	2,368	322	1,680	73	4	122	167
Mont.	4	4,490	3,336	828	23	....	139	164	3,562	660	2,517	94	1	151	139	212	57	108	12	1	6	28
Nebr.	7	18,280	12,436	4,241	100	7	685	811	11,770	1,677	9,045	177	1	418	451	862	157	594	31	2	18	60
Nev.	3	1,003	658	272	....	....	36	37	955	163	716	2	....	42	32	36	9	22	....	....	4	1
N. H.	4	9,207	7,594	1,072	21	....	253	357	2,737	479	1,984	51	1	114	108	372	94	236	8	....	18	24
N. J.	16	58,677	45,361	8,625	251	17	2,383	2,040	27,631	5,495	18,642	1,032	14	1,548	900	2,444	442	1,635	89	1	104	173
N. M.	3	1,625	1,003	444	7	1	80	90	1,662	212	1,290	24	....	70	66	54	11	33	1	2	2	....
N. Y.	47	162,260	114,574	33,052	805	45	7,125	6,659	139,277	18,241	99,938	4,101	141	10,604	6,252	14,656	2,106	10,414	303	23	620	1,141
N. C.	13	6,113	3,532	1,656	33	5	580	307	16,324	820	13,778	119	6	946	655	35	5	20	3	....	4	....
N. Dak.	4	4,250	2,787	1,157	15	1	108	182	3,666	694	2,679	30	2	97	164	1,111	192	743	32	5	29	110
Ohio	26	77,596	58,232	13,391	420	66	2,747	3,040	50,778	9,465	35,864	1,315	38	2,454	1,642	8,156	1,580	5,389	249	14	375	549
Okla.	11	14,442	8,393	4,260	29	3	1,050	707	15,075	1,289	12,389	53	2	687	655	217	36	143	10	....	9	19
Ore.	5	16,682	8,593	2,014	72	6	521	541	10,951	1,966	7,666	298	7	567	447	655	90	313	26	2	30	54
Pa.	36	119,086	86,433	20,097	543	115	6,461	5,437	81,114	14,502	56,082	1,340	55	5,733	3,402	7,507	1,121	5,089	187	11	467	632
R. I.	4	10,401	8,165	1,269	32	5	511	419	3,489	600	2,470	90	....	208	121	794	148	545	12	3	51	35
S. C.	8	1,247	216	698	2	....	300	71	7,105	101	5,943	6	6	701	348	20	2	11	1	....	2	....
S. Dak.	4	8,483	5,712	2,096	42	14	248	371	4,507	859	3,314	46	5	125	158	770	122	539	20	10	20	9
Tenn.	11	9,383	5,285	2,354	29	31	1,178	506	19,829	1,419	15,510	128	33	1,938	801	100	14	63	2	....	12	....
Texas	23	15,341	6,302	6,774	43	3	1,559	660	37,501	1,860	31,262	149	5	2,668	1,557	558	59	417	13	1	28	4
Utah	4	4,067	2,906	851	21	1	155	133	5,318	954	3,935	69	8	189	163	119	30	65	8	....	5	1
Vt.	3	7,241	5,829	822	20	2	239	329	2,458	498	1,756	37	1	84	83	174	48	90	2	....	18	16
Va.	11	10,223	5,696	2,848	57	18	1,194	410	16,783	1,121	13,346	141	14	1,517	644	74	17	37	4	....	8	8

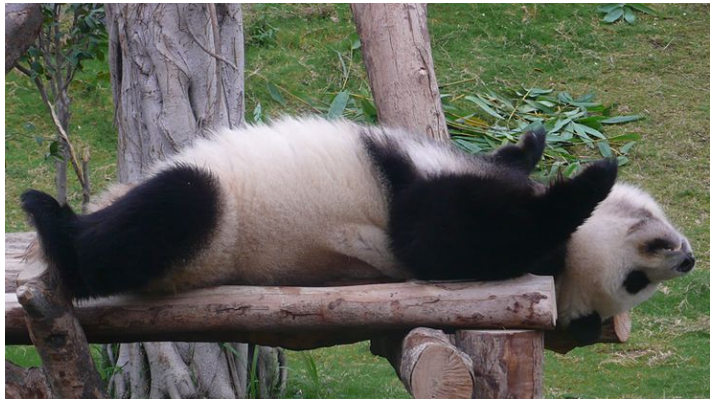
Demo

lec09.ipynb

CC BY NC SA

1760929

2-minute stretch break!

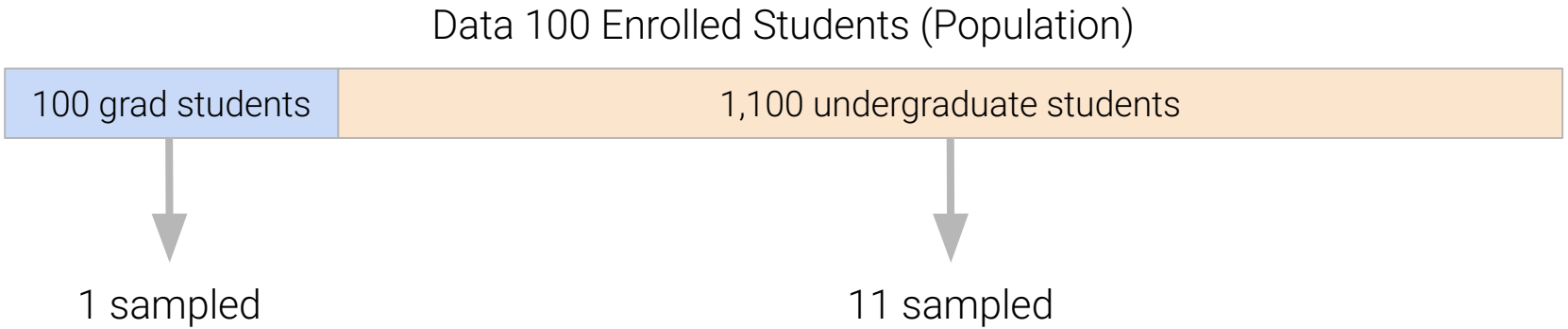


Another Example Sampling Scheme

1760929

I want to interview a representative sample of **12 students** enrolled in Data 100.

- Suppose there are **1,200 students** in Data 100.
- **100 students** are graduate students. The remaining **1,100** are undergraduates.
- I conduct an SRS with **n=1** on the 100 graduate students, and an SRS with **n=11** on the 1,100 undergraduates.



CC BY NC SA

1760929

Does every student have the same probability of being selected for an interview?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Another Example Sampling Scheme

1760929

Data 100 Enrolled Students (Population)

1 sampled

11 sampled

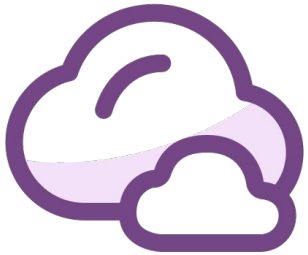
1. Does each student have the same probability of being selected?

Yes.

Each student is chosen with probability 1/100.

CC BY NC SA

1760929



Is there any benefit or downside to sampling this way, instead of selecting 12 students at random from all 1200 enrolled students?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

slido

Another Example Sampling Scheme

1760929

Data 100 Enrolled Students (Population)

1 sampled

11 sampled

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

Yes.

Each student is chosen with probability 1/100.

2. Is there any benefit or downside to sampling this way?

Yes, a benefit!

We have **guaranteed proportional representation** of undergrads and grad students.

In other words, we have **reduced chance error** (i.e., variance).

CC BY NC SA

Stratified Random Sampling

1760929

Data 100 Enrolled Students (Population)

Stratum 1
1 sampled

Stratum 2
11 sampled

Sampling frame is divided into non-overlapping **strata** according to chosen characteristics.

- Then, a simple random sample (SRS) is conducted on each **stratum**, with each sample size proportional to the stratum size.

Note: In this example, the sampling frame perfectly overlaps with the population.

CC BY NC SA

Stratified Random Sampling

1760929

Data 100 Enrolled Students (Population)

Stratum 1
1 sampled

Stratum 2
11 sampled

Benefits

- Guaranteed proportional representation from groups of interest.
- Reduced chance error (i.e., variance). Less likely to get an unrepresentative sample.

Limitations

- Adds a layer of complexity to data analysis.
- Minimal chance error reduction with large sample size and "big enough" strata.
- Population proportions of group characteristics not always known (e.g., if we did not know # of undergrads and grads in Data 100)

CC BY NC SA

1760929

The best way to get a representative sample is through *randomization*.

(Though, sometimes it's not easy or possible.)

What can we do with a non-representative sample?

CC BY NC SA

Non-representative Samples Are Common

1760929

We often cannot obtain a truly representative sample.

- Customer surveys → Not everyone responds or provides contact info.
- Election polling → Not all voters are reachable or want to talk to you.
- Clinical trials → Patients have to voluntarily join a trial.

CC BY NC SA

Non-representative Sampling Scheme

1760929

Data 100 Enrolled Students (**Population**)

100 grad students	1,100 undergraduate students

Twenty Data 100 students in a room (Non-representative **Sample**)

10 grad students	10 undergraduate students
9 are enjoying Data 100 this week	**5** are enjoying Data 100 this week

Based on just these 20 students, what's your best guess of the percentage of **all** Data 100 students who are enjoying Data 100 this week?

CC BY NC SA

1760929

Based on just these 20 students, what's your best guess of the percentage of all Data 100 students who are enjoying Data 100?

Presenting with animations, GIFs or speaker notes? Enable our Chrome extension

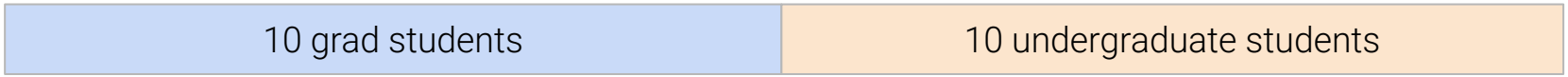
slido

Non-representative Sampling Scheme

1760929

Data 100 Enrolled Students (**Population**)

Twenty Data 100 students in a room (Non-representative **Sample**)



9 are enjoying Data 100 this week

5 are enjoying Data 100 this week

There are a lot more undergraduates in Data 100 than grad students.

- **Overweight** the opinions of the 10 undergraduates relative to the 10 grad students.

To trust our estimate, we might **assume** that the 10 grad students are representative of all 100 grad students, and the 10 undergrads of all 1100 undergraduates. Giant assumption!

CC BY NC SA

Non-representative Sampling Scheme

1760929

Data 100 Enrolled Students (**Population**)

100 grad students	1,100 undergraduate students

Twenty Data 100 students in a room (Non-representative **Sample**)

10 grad students	10 undergraduate students
9 are enjoying Data 100 this week	5 are enjoying Data 100 this week

(9 / 10) * (100 / 1200) + (5 / 10) * (1100 / 1200) = **53.3%**

Post-stratification: <u>After</u> sampling, use knowledge about the population to **reweight** responses.

CC BY NC SA

Post-stratification: Definition

1760929

1. Divide your sample and population into distinct **cells** according to chosen characteristics (e.g., undergrad and grad).
2. Calculate the **overall** response in each sample cell (e.g., proportion enjoying Data 100)
3. Aggregate over the sample cells, proportionally **weighting** each sample cell by the **size of the corresponding population cell**.

Assumptions:

1. The population cell sizes are accurate.
2. Each sample cell is representative of the corresponding population cell. This is a big assumption!

What if you don't have representation in the sample for every population cell? See MRP.

CC BY NC SA

Post-stratification of 1936 Literary Digest poll results

1760929

1. Divide your sample and population into distinct **cells** according to chosen characteristics (e.g., U.S. state of residence and political party).
2. Calculate the **overall** response in each sample cell (e.g., number of intended votes).
3. Aggregate over the sample cells, proportionally **weighting** each sample cell by the **size of the corresponding population cell**.

1936 Literary Digest Poll (Sample)

	Roosevelt (Democrat)	**Landon (Republican)**
Alabama	10,082	3,060
Arizona	1,975	2,337
...	...	...
Wyoming	1,533	2,526

Actual 1936 Votes (Population)

	Roosevelt (Democrat)	**Landon (Republican)**
Alabama	?	?
Arizona	?	?
...	...	...
Wyoming	?	?

From the perspective of Literary Digest in 1936, we don't know the population cell sizes.

CC BY NC SA

Respondents Reported Their Actual 1932 Votes!

1760929

Final Report "Literary Digest" 1936 Presidential Poll

	Electoral Vote	Landon 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Roosevelt 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Lemke 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated
Ala.	11	3,060	1,218	1,298	3	3	412	126	10,082	371	8,530	50	1	736	394	68	5	49	4	.....	4	6
Ariz.	3	2,337	1,431	647	18	.....	129	112	1,975	248	1,555	33	.....	70	69	104	22	52	8	.....	10	12
Ark.	9	2,724	1,338	953	7	9	274	143	7,608	228	6,655	16	8	373	328	138	14	98	4	3	9	10
Calif.	22	89,516	65,360	16,200	315	53	3,519	4,069	77,245	15,165	53,520	1,816	63	3,578	3,103	4,977	1,620	2,560	117	25	163	492
Colo.	6	15,949	11,872	2,714	131	12	637	583	10,025	1,747	7,256	284	13	439	286	579	136	333	29	2	26	53
Conn.	8	28,809	22,939	3,376	111	7	1,230	1,146	13,413	2,584	9,113	408	6	788	514	1,489	245	1,006	53	3	70	112
Del.	3	2,918	2,343	328	9	.....	134	104	2,048	503	1,345	34	.....	96	70	35	6	19	3	.....	2	5
Fla.	7	6,087	3,121	2,051	13	5	594	303	8,620	635	6,924	41	.....	614	406	195	37	116	6	2	12	22
Ga.	12	3,948	1,239	1,817	5	11	708	168	12,915	379	10,377	42	9	1,569	539	35	3	23	1	.....	6	2
Idaho	4	3,653	2,672	698	9	8	103	163	2,611	398	1,989	30	8	89	97	224	69	109	8	11	9	18
Ill.	29	123,297	85,112	25,885	573	69	6,506	5,152	79,035	14,793	54,612	1,542	57	4,790	3,241	6,415	1,172	4,219	169	17	304	534
Ind.	14	42,805	31,913	7,644	134	49	1,290	1,775	26,663	4,513	20,247	302	22	719	860	2,166	476	1,352	64	11	73	190
Iowa	11	31,871	22,823	6,164	135	26	1,272	1,451	18,614	3,190	13,611	258	14	829	712	2,829	560	1,831	86	11	88	253
Kans.	9	35,408	25,315	6,489	147	15	1,466	1,976	20,254	4,182	14,121	257	11	846	837	902	226	482	52	1	43	98
Ky.	11	13,365	8,957	2,939	35	14	793	627	16,592	1,586	13,594	95	6	703	608	732	69	554	24	.....	31	54
La.	10	3,686	1,366	1,742	9	3	384	182	7,902	445	6,401	39	.....	697	320	841	35	667	23	2	55	59
Maine	5	11,742	8,619	1,567	25	35	713	783	5,337	635	3,820	41	1	289	551	418	64	277	3	2	42	30
Md.	8	17,463	9,754	4,685	110	2	1,479	1,433	18,341	1,891	13,540	328	5	1,366	1,211	614	56	422	22	1	34	79
Mass.	17	87,449	70,567	10,105	330	31	3,213	3,203	25,965	5,141	17,499	744	16	1,635	930	5,415	1,002	3,670	133	3	236	371
Mich.	19	51,478	38,526	8,665	287	22	2,113	1,865	25,686	5,114	17,402	748	26	1,472	924	3,376	680	2,145	128	4	130	239
Minn.	11	30,762	22,386	5,958	109	3	972	1,334	20,733	3,699	14,855	511	22	861	785	5,426	804	3,893	115	14	157	443
Miss.	9	848	269	394	1	.....	137	47	6,080	88	5,396	8	1	298	289	43	5	32	1	.....	2	3
Mo.	15	50,022	33,551	11,149	244	45	2,975	2,058	38,267	4,463	30,608	455	15	1,485	1,241	2,368	322	1,680	73	4	122	167
Mont.	4	4,490	3,336	828	23	.....	139	164	3,562	660	2,517	94	1	151	139	212	57	108	12	1	6	28
Nebr.	7	18,280	12,436	4,241	100	7	685	811	11,770	1,677	9,045	177	2	418	451	862	157	594	31	2	18	60
Nev.	3	1,003	658	272	.....	.....	36	37	955	163	716	2	.....	42	32	36	9	22	.....	.....	4	1
N. H.	4	9,207	7,504	1,072	21	.....	253	357	2,737	479	1,984	51	1	114	108	372	84	238	8	.....	18	24
N. J.	16	58,677	45,361	8,625	251	17	2,383	2,040	27,631	5,495	18,642	1,032	14	1,548	900	2,444	442	1,633	89	1	104	175
N. M.	3	1,625	1,003	444	7	1	80	90	1,662	212	1,290	24	.....	70	66	54	13	33	1	2	2	3
N. Y.	47	162,260	114,574	33,052	805	45	7,125	6,659	139,277	18,241	99,938	4,101	141	10,604	6,252	14,656	2,106	10,414	303	20	670	1,143
N. C.	13	6,113	3,532	1,656	33	5	580	307	16,324	820	13,778	119	6	946	655	35	5	20	3	.....	4	3
N. Dak.	4	4,250	2,787	1,157	15	1	108	182	3,666	694	2,679	30	2	97	164	1,111	192	743	32	5	29	110
Ohio	26	77,896	58,232	13,391	420	66	2,747	3,040	50,778	9,465	35,864	1,315	38	2,454	1,642	8,156	1,580	5,389	249	14	375	549
Okla.	11	14,442	8,393	4,260	29	3	1,050	707	15,075	1,289	12,389	53	2	687	655	217	36	143	10	.....	9	19
Ore.	5	11,747	8,593	2,014	72	6	521	541	10,951	1,966	7,666	298	7	567	447	655	196	313	46	7	30	6
Pa.	36	119,086	86,433	20,097	543	115	6,461	5,437	81,114	14,502	56,082	1,340	55	5,733	3,402	7,507	1,121	5,089	187	11	467	632
R. I.	4	10,401	8,165	1,269	32	5	511	419	3,489	600	2,470	90	.....	208	121	794	148	545	12	3	31	55
S. C.	8	1,247	216	658	2	.....	300	71	7,105	101	5,943	6	6	701	348	20	2	11	1	.....	2	.....
S. Dak.	4	8,483	5,712	2,096	42	14	248	371	4,507	859	3,314	46	5	125	158	770	122	539	20	10	20	59
Tenn.	11	9,883	5,785	2,354	29	31	1,178	506	19,829	1,419	15,510	128	33	1,938	801	100	14	63	2	.....	12	9
Texas	23	15,341	6,302	6,774	43	3	1,559	660	37,501	1,860	31,262	149	5	2,668	1,557	558	58	417	13	1	28	43
Utah	4	4,067	2,906	851	21	1	155	133	5,318	954	3,935	69	8	189	163	119	30	65	8	.....	5	11
Vt.	3	7,241	5,829	822	20	2	239	329	2,458	498	1,756	37	.....	84	83	174	48	90	2	.....	18	16
Va.	11	10,223	5,696	2,848	57	18	1,194	410	16,783	1,121	13,346	141	14	1,517	644	74	17	37	4	.....	8	8

CC BY NC SA

Post-stratification of 1936 Literary Digest poll results

1760929

What if we calculate weights with **1932** reported votes and **1932** election results, and then apply to 1936?

Reported **1932** Votes from 1936 poll (Sample)

	Democrat	Republican
Alabama	9,828	1,589
Arizona	2,202	1,679
…	…	…
Wyoming	1,654	2,072

33x more Democrats in election than poll

19x more Republicans in election than poll

Actual **1932** Votes (Population)

	Democrat	Republican
Alabama	207,910	34,675
Arizona	79,294	36,104
…	…	…
Wyoming	54,370	39,583

Democrats are underrepresented in the Wyoming poll; Republicans are overrepresented.

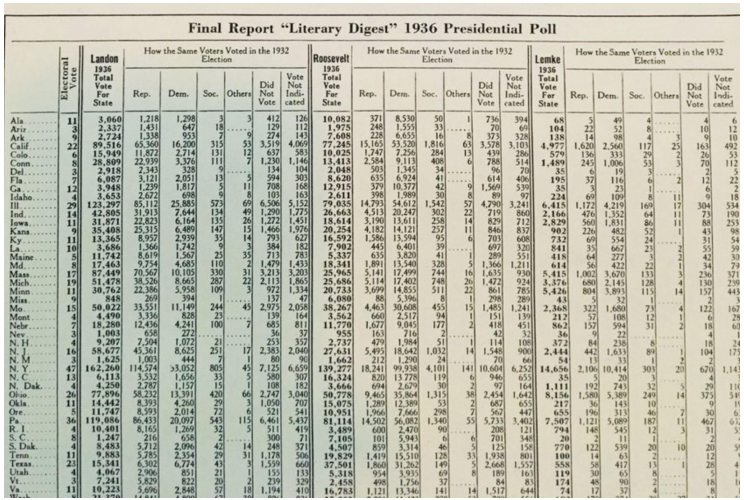
Revisiting the 1936 Literary Digest Poll

1760929

The 1936 Literary Digest poll predicted Roosevelt would lose the presidential election with 43% of the vote.

- In reality, Roosevelt won with 61% of the vote.

By using post-stratification, could Literary Digest have made a more accurate prediction?



Final Report "Literary Digest" 1936 Presidential Poll

	Electoral Vote	Landon 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Roosevelt 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated	Lemke 1936 Total Vote For State	How the Same Voters Voted in the 1932 Election: Rep.	Dem.	Soc.	Others	Did Not Vote	Vote Not Indicated
Ala.	11	3,060	1,218	1,298	3	3	412	126	10,082	371	8,530	50	1	736	394	68	5	49	4	....	4	6
Ariz.	3	2,337	1,431	647	18	....	129	112	1,975	248	1,555	33	....	70	69	104	22	52	8	....	10	12
Ark.	9	2,724	1,338	953	7	9	274	143	7,608	228	6,655	16	8	373	328	138	14	98	4	5	9	10
Calif.	22	89,516	65,360	16,200	315	53	3,519	4,069	77,245	15,165	53,520	1,816	63	5,578	3,103	4,977	1,620	2,560	117	25	163	492
Colo.	6	15,949	11,872	2,714	131	12	637	583	10,025	1,747	7,256	284	13	439	286	579	136	333	29	2	26	53
Conn.	8	28,809	22,939	3,376	111	7	1,230	1,146	13,413	2,584	9,113	408	6	788	514	1,489	245	1,006	53	3	70	112
Del.	3	2,918	2,343	328	9	....	134	104	2,048	503	1,345	34	....	96	70	35	6	19	3	....	2	5
Fla.	7	6,087	3,121	2,051	13	5	594	303	8,620	635	6,924	41	....	614	406	195	37	116	6	2	12	22
Ga.	12	3,948	1,239	1,817	5	11	708	168	12,915	329	10,372	42	9	1,569	539	35	3	23	1	....	6	2
Idaho	4	3,653	2,672	698	9	8	103	163	2,611	398	1,989	30	8	89	97	224	69	109	8	11	9	18
Ill.	29	123,297	85,112	25,885	573	69	6,506	5,152	79,035	14,793	54,612	1,542	57	4,790	3,241	6,415	1,172	4,219	169	17	304	534
Ind.	14	42,505	31,913	7,644	154	49	1,200	1,775	26,663	4,513	20,247	302	22	719	860	2,166	476	1,352	64	11	73	190
Iowa.	11	31,871	22,823	6,164	135	26	1,272	1,451	18,614	3,190	13,611	258	14	829	712	2,829	560	1,831	86	11	88	253
Kans.	9	35,408	25,315	6,409	147	15	1,466	1,976	20,254	4,182	14,121	257	11	846	837	902	226	482	52	1	43	98
Ky.	11	13,365	8,957	2,939	35	14	793	627	16,592	1,586	13,594	95	6	703	608	732	69	554	24	....	31	54
La.	10	3,686	1,366	1,742	9	3	384	182	7,902	445	6,401	39	....	697	320	841	35	662	23	2	55	59
Maine	5	11,742	8,619	1,567	25	35	713	783	5,337	635	3,829	41	1	289	551	418	64	272	3	2	42	30
Md.	8	17,463	9,754	4,685	110	2	1,479	1,433	18,341	1,891	13,540	328	5	1,366	1,211	614	56	422	22	1	34	79
Mass.	17	87,449	70,567	10,105	330	31	3,213	3,203	25,965	5,141	17,499	744	16	1,635	930	5,415	1,002	3,670	133	3	236	371
Mich.	19	51,478	38,526	8,665	287	22	2,113	1,865	25,686	5,114	17,402	748	26	1,472	924	3,376	680	2,145	128	4	130	299
Minn.	11	30,762	22,386	5,958	109	3	972	1,334	20,733	3,699	14,855	511	22	861	785	5,426	804	3,893	115	14	157	443
Miss.	9	848	269	394	1	....	157	47	6,000	88	5,396	8	1	298	209	43	5	32	1	....	3	2
Mo.	15	50,022	33,551	11,149	244	45	2,975	2,058	38,267	4,463	30,608	455	15	1,485	1,241	2,368	322	1,680	73	4	122	167
Mont.	4	4,490	3,336	828	23	....	139	164	3,562	660	2,517	94	1	151	139	212	57	108	12	1	6	28
Nebr.	7	18,280	12,436	4,241	100	7	685	811	11,770	1,677	9,045	177	1	418	451	862	157	594	31	2	18	60
Nev.	3	1,003	658	272	....	....	36	37	955	163	716	2	....	42	32	36	9	22	....	....	4	1
N. H.	4	9,207	7,594	1,072	21	....	253	357	2,737	479	1,984	51	1	114	108	372	94	236	8	....	18	16
N. J.	16	58,677	45,361	8,625	251	17	2,383	2,040	27,631	5,495	18,642	1,032	14	1,548	900	2,444	442	1,635	89	1	104	173
N. M.	3	1,625	1,003	444	7	1	80	90	1,662	212	1,290	24	....	70	66	54	11	33	1	2	2	5
N. Y.	47	162,260	114,574	33,052	805	45	7,125	6,659	139,277	18,241	99,938	4,101	141	10,604	6,252	14,656	2,106	10,414	303	23	620	1,141
N. C.	13	6,113	3,532	1,656	33	5	580	307	16,324	820	13,778	119	6	946	655	35	5	20	3	....	4	....
N. Dak.	4	4,250	2,787	1,157	15	1	108	182	3,666	694	2,679	30	2	97	164	1,111	192	743	32	5	29	110
Ohio	26	77,596	58,232	13,391	420	66	2,747	3,040	50,778	9,465	35,864	1,315	38	2,454	1,642	8,356	1,580	5,389	249	14	375	549
Okla.	11	14,442	8,393	4,260	29	3	1,050	707	15,075	1,289	12,389	53	2	687	655	217	36	143	10	....	9	19
Ore.	7	16,682	8,593	2,014	72	6	521	541	10,951	1,966	7,666	298	7	567	447	655	102	313	26	2	30	82
Pa.	36	119,086	86,433	20,097	543	115	6,461	5,437	81,114	14,502	56,082	1,340	55	5,733	3,402	7,507	1,121	5,089	187	11	467	632
R. I.	4	10,401	8,165	1,269	32	5	511	419	3,489	600	2,470	90	....	208	121	794	148	545	12	3	31	55
S. C.	8	1,247	216	698	2	....	300	71	7,105	101	5,943	6	6	701	348	20	2	11	1	....	2	4
S. Dak.	4	8,483	5,712	2,096	42	14	248	371	4,507	859	3,314	46	5	125	158	770	122	539	20	10	20	59
Tenn.	11	9,883	5,285	2,354	29	31	1,178	506	19,829	1,419	15,510	128	33	1,938	801	100	14	63	2	....	12	9
Texas	23	15,341	6,302	6,774	43	3	1,559	660	37,501	1,860	31,262	149	5	2,668	1,557	558	59	417	13	1	28	40
Utah	4	4,067	2,906	851	21	1	155	133	5,318	954	3,935	69	8	189	163	119	30	65	8	....	5	1
Vt.	3	7,241	5,829	822	20	2	239	329	2,458	498	1,756	37	1	84	83	174	48	90	2	....	18	16
Va.	11	10,223	5,696	2,848	57	18	1,194	410	16,783	1,121	13,346	141	14	1,517	644	74	12	37	4	....	8	13

Demo

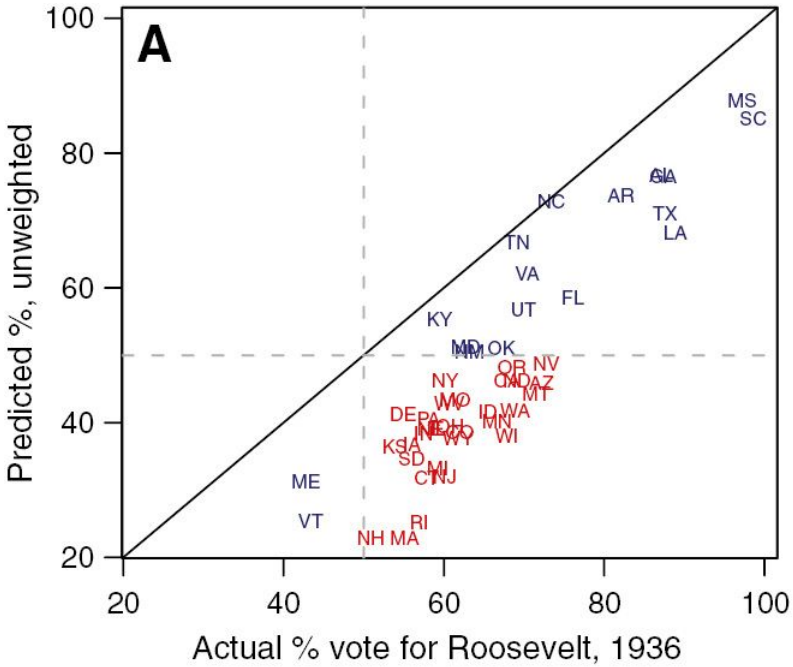
lec09.ipynb

CC BY NC SA

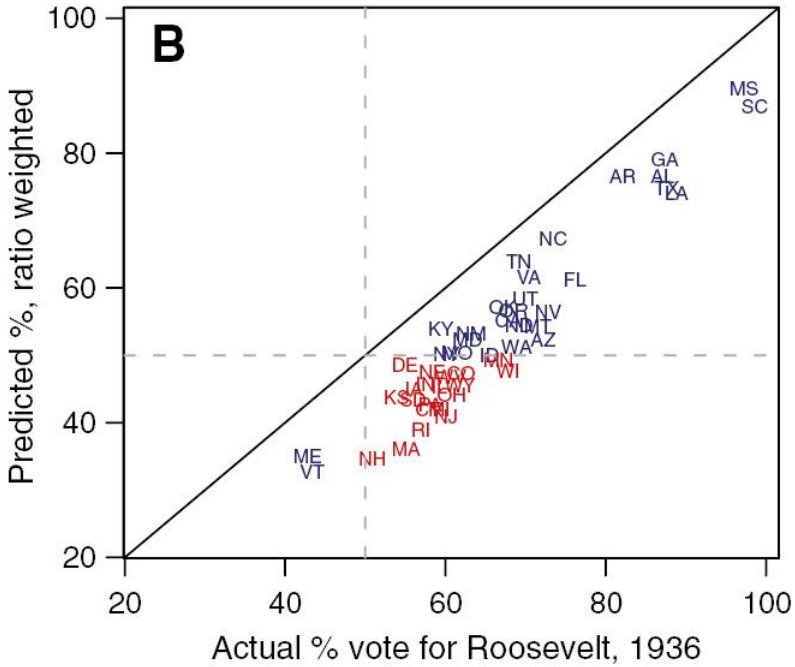
Post-stratification and the 1936 Literary Digest Poll

1760929

Raw Results (Unweighted)
Prediction: Landon wins easily.



Post-stratified Results (Weighted)
Prediction: Roosevelt wins by a small margin.

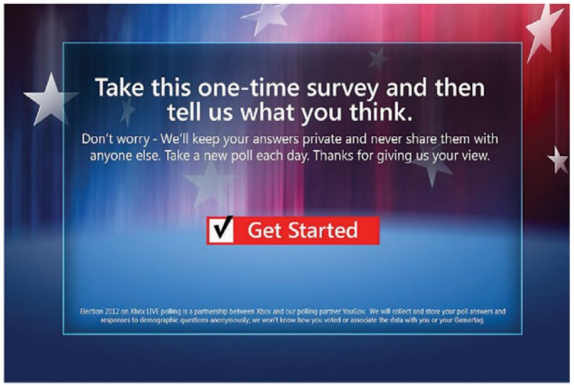


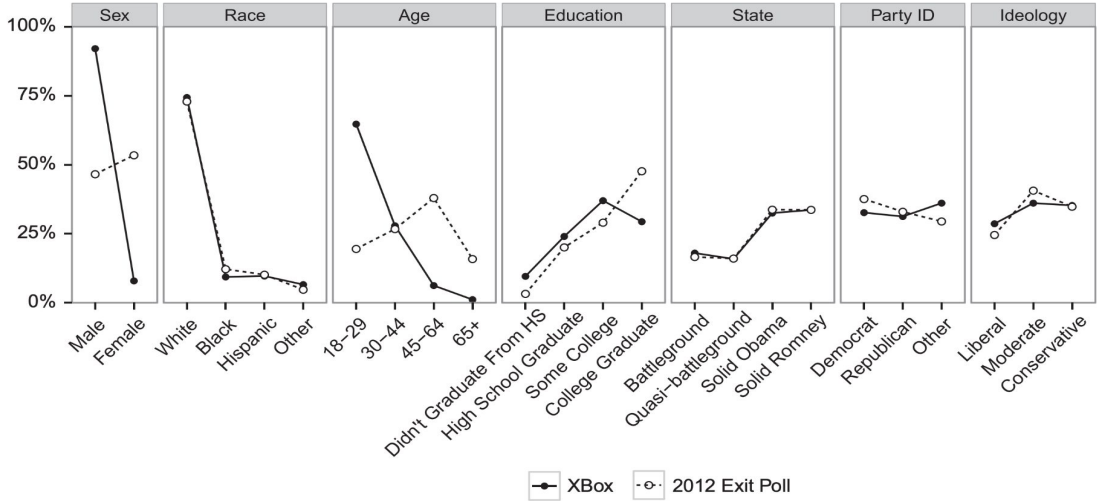
Lohr and Brick (2017)

CC BY NC SA

XBox Polling: 2012 Presidential Election

1760929

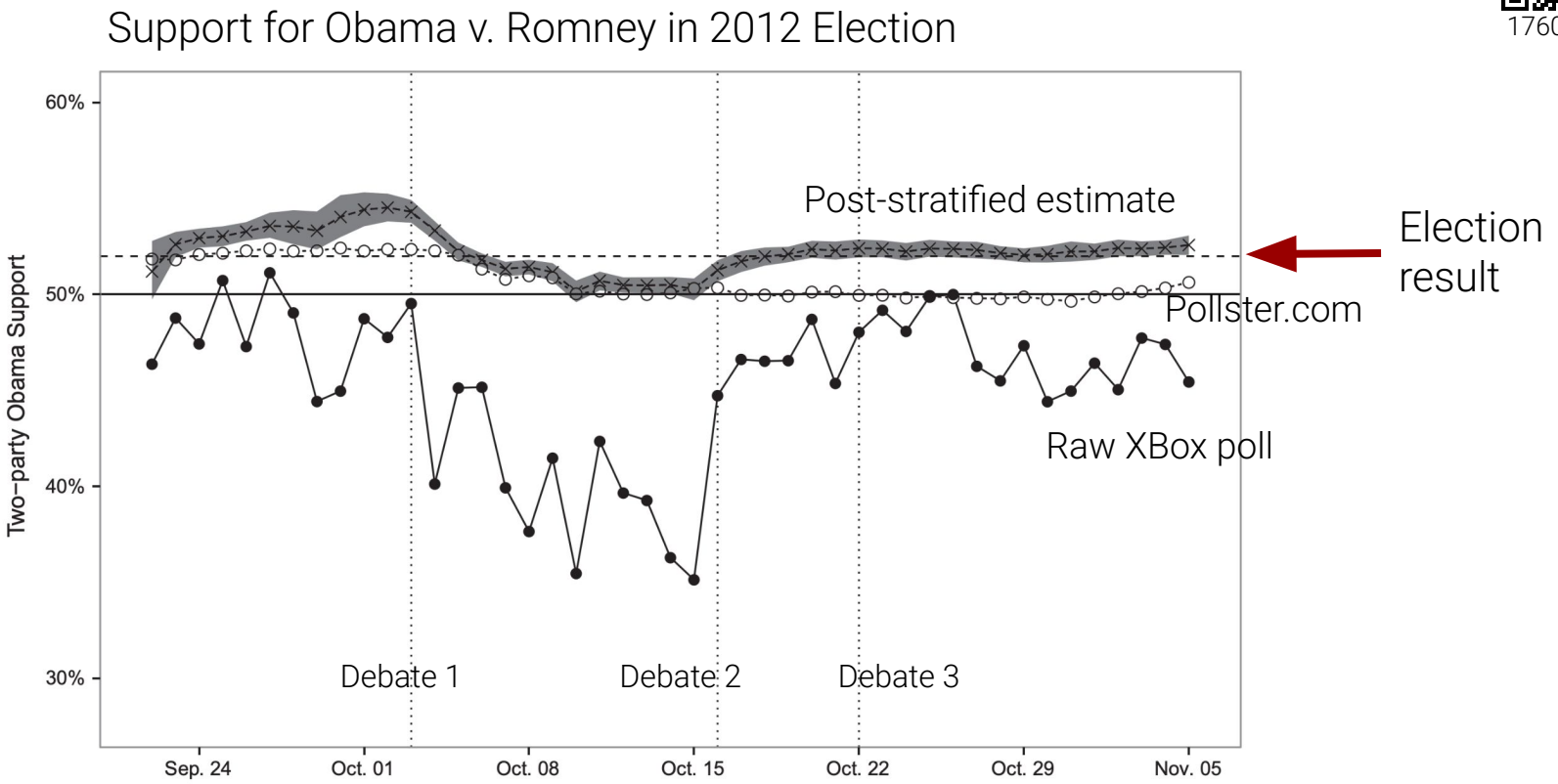




Wang et al. (2014)

XBox Polling: 2012 Presidential Election

1760929



Wang et al. (2014)

CC BY NC SA

1760929

We made it!

Lecture 9, Data 100 Spring 2025

- Censuses and Surveys
- Sampling: A Case Study
- Sampling Errors
- Types of Sampling
- Post-stratification

CC BY NC SA

1760929

LECTURE 9

Sampling

Content credit: Acknowledgments

CC BY NC SA