



DS-UA 112

Introduction to Data Science

Week 2: Lecture 2

Collecting Data - Approaches to Sampling





How can we collect data?
How can a dataset be non-
representative?

DS-UA 112

Introduction to Data Science

Week 2: Lecture 2

Collecting Data - Approaches to Sampling

Adapted from Nolan, Hug, and Salganik



Announcements

- ▶ Please check Week 2 agenda on NYU Classes

- ▶ Homework 1

- ▶ Lab 2

- ▶ Survey 1

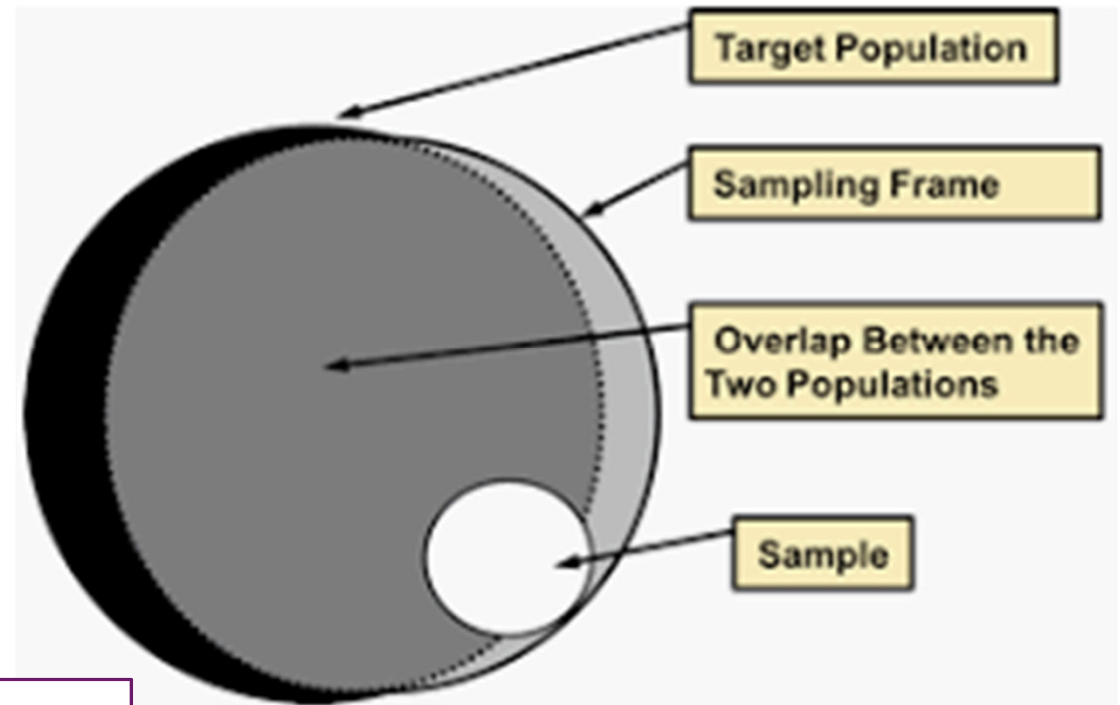
- ▶ Remember to post to Piazza

Check the Calendar
linked to NYU Classes
for important dates

applied
algorithm don't
understanding deep
field
clean
idea
skill job
code
world
large
hope
method
analyze
help
library
create expand actual
practical
real
python
application
science
data
work
good
project
knowledge
basic
class
making
experience
gain
hand
tool
world
code
large
hope
method
analyze
help
library
create expand actual

Review

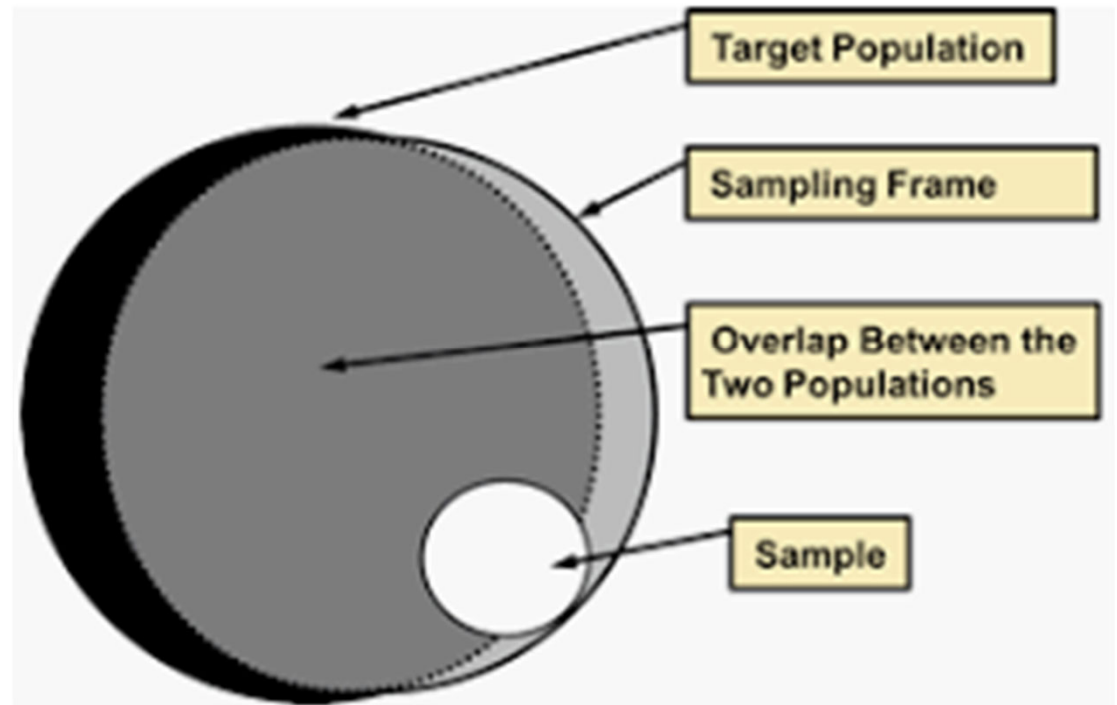
- ▶ Collecting Data
 - ▶ Sample
 - ▶ Sampling Frame
 - ▶ Population
- ▶ Representative Data
 - ▶ Census
 - ▶ Administrative Dataset



Representative of the
population or
sampling frame.

Review

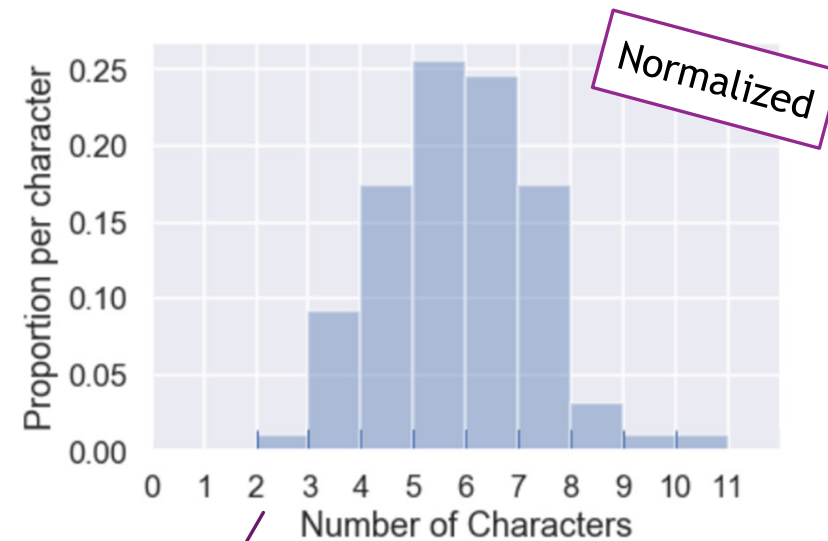
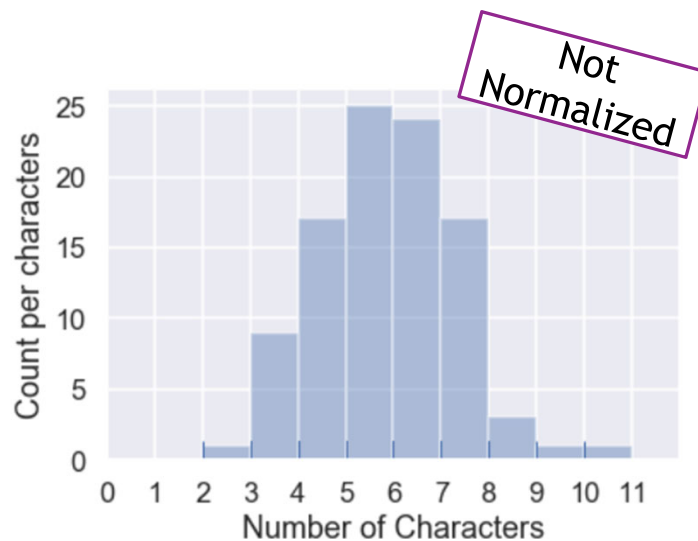
- ▶ Non-Representative Data
 - ▶ Self Selected Sample
 - ▶ Judgement Sample
 - ▶ Convenience Sample
- ▶ Random Samples
 - ▶ Simple Random Sample



Non-Representative may
or may not be **biased**.

Review

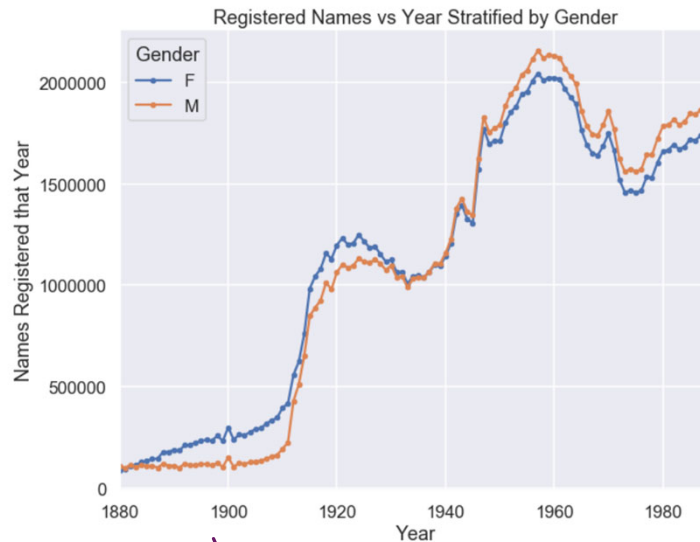
- Visualize to Quantify
 - Histogram
 - Displays information about **one** variable
 - Line Chart
 - Displays information about **two** variables



Ranges are called bins or buckets

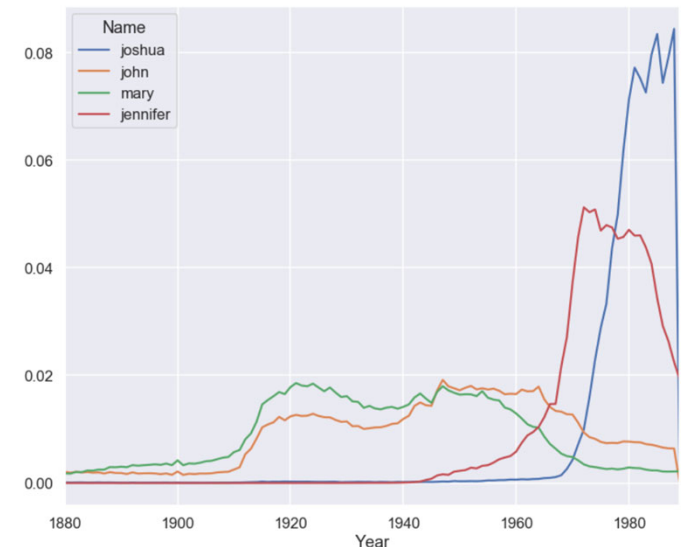
Review

- Visualize to Quantify
 - Histogram
 - Displays information about **one** variable
 - Line Chart
 - Displays information about **two** variables



Each subset is called a **stratum**

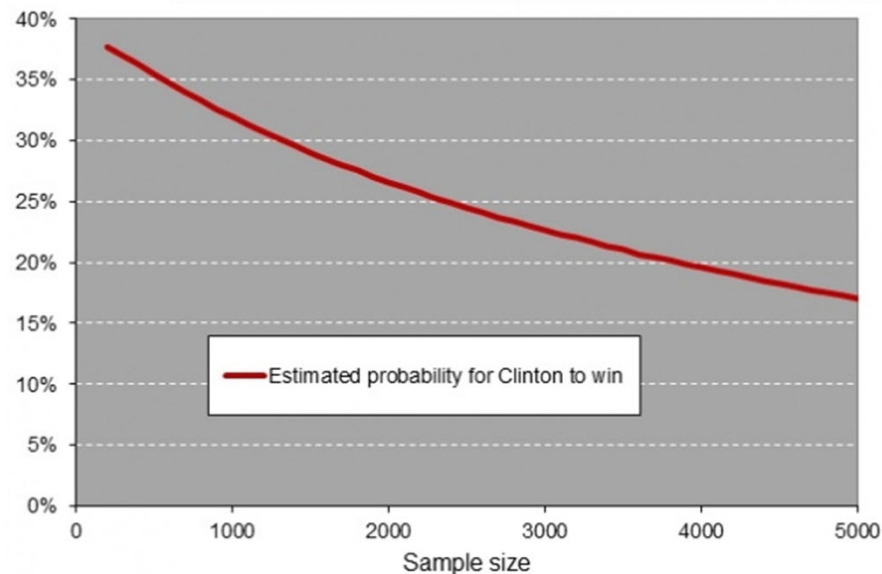
Stratify means
split set into
subsets



Election Polling

- ▶ AAPOR attributes incorrect predictions to
 - ▶ change in voters' preferences just before the election
 - ▶ overrepresentation of college graduates in some poll samples
 - ▶ Trump voters revealing their preferences later in race

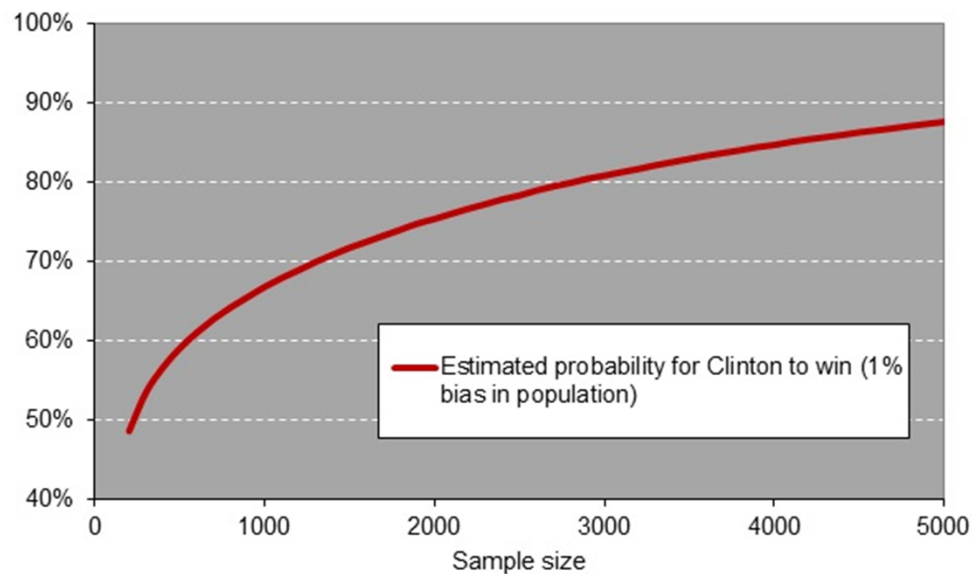
"The point is that the polls very likely sampled from a population that just was not sufficiently representing Republicans"



Election Polling

- ▶ AAPOR attributes incorrect predictions to
 - ▶ change in voters' preferences just before the election
 - ▶ overrepresentation of college graduates in some poll samples
 - ▶ Trump voters revealing their preferences later in race

"The point is that the polls very likely sampled from a population that just was not sufficiently representing Republicans"



Agenda

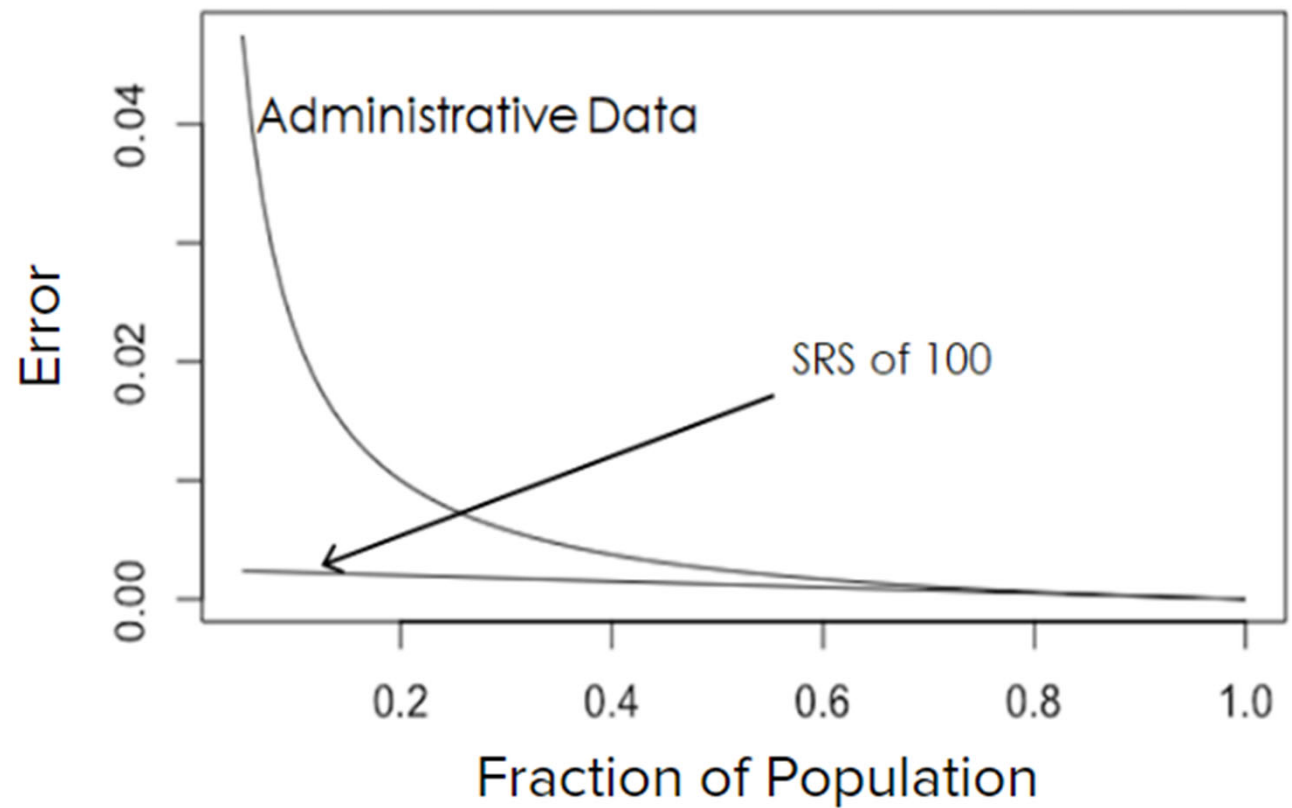
- ▶ Random Sampling
 - ▶ SRS, Cluster, Stratified
- ▶ Probability
 - ▶ Addition, Multiplication, Complement Rules
- ▶ Summarize to Quantify
 - ▶ Expectation
 - ▶ Bias

References

- ▶ Nolan, Lau, Gonzalez (Chapter 2)
 - ▶ <https://cp71.github.io/textbook>
- ▶ Salganik (Chapter 3)

Randomization

- Randomization can produce representative datasets from few observations



Exercise

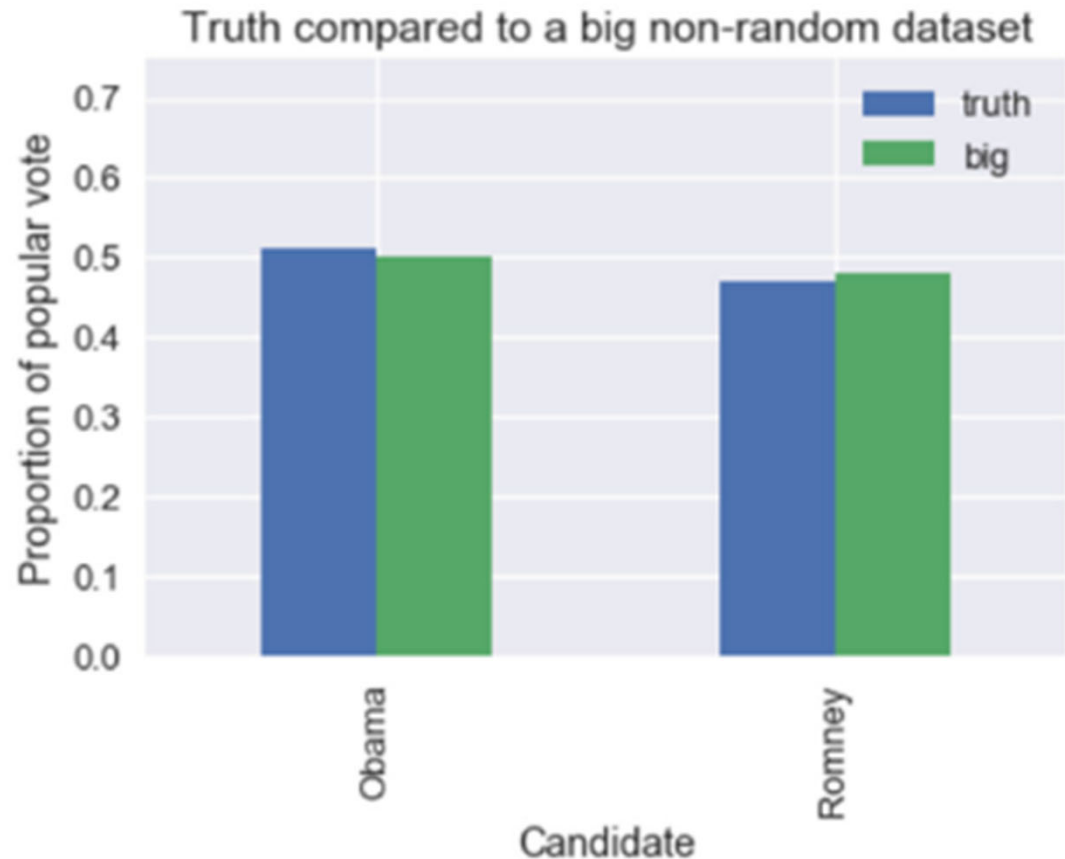
What does the diagram look like for Simple Random Sampling?

- ▶ Simple Random Sample
- ▶ Cluster Sample
- ▶ Stratified Sample

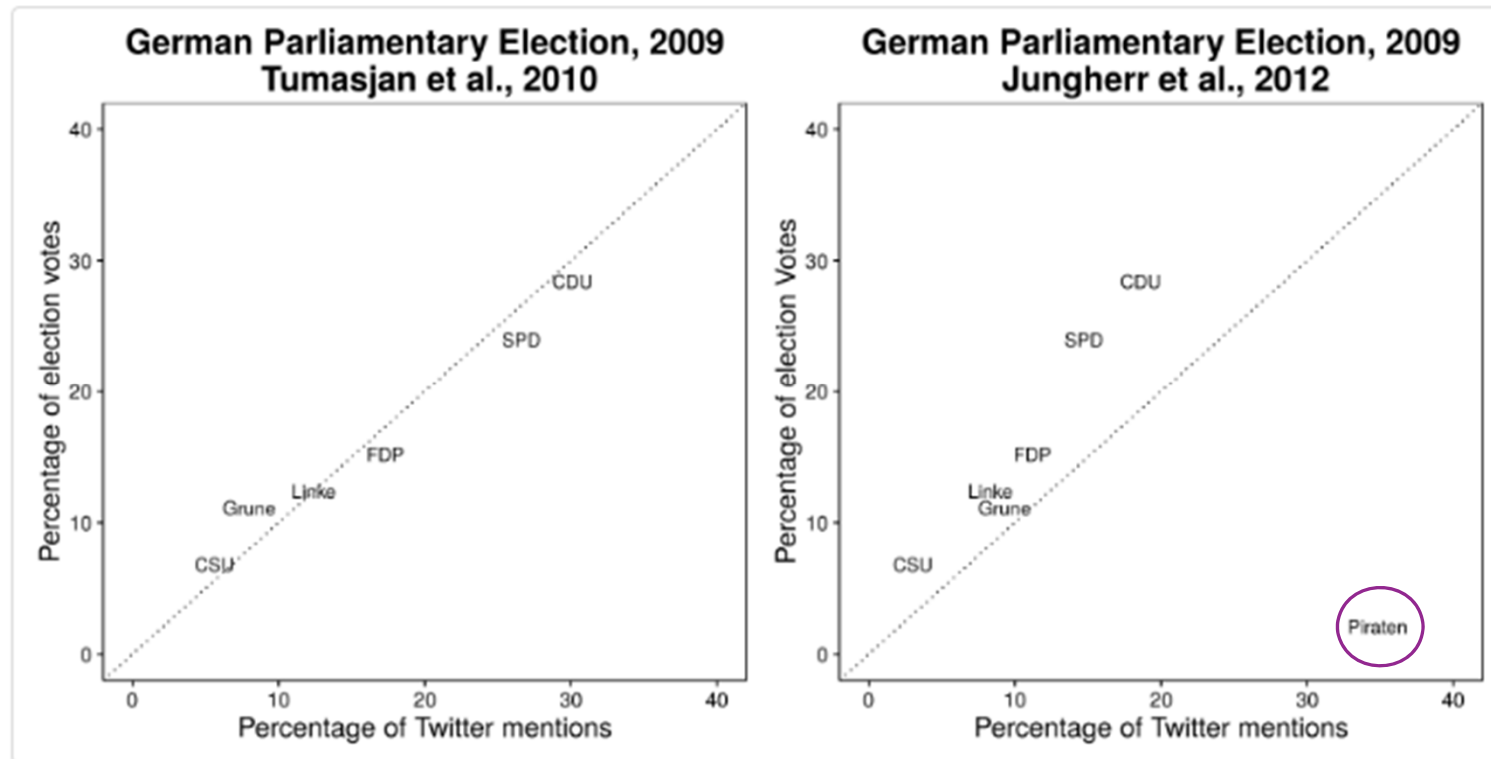


Randomization

- Simulations shows that large datasets might be less accurate than small datasets in the presence of bias



Judgement Sample



Self Selected Sample

WHY THE 1936 *LITERARY DIGEST* POLL FAILED


PEVERILL SQUIRE

Abstract The *Literary Digest* poll of 1936 holds an infamous place in the history of survey research. Despite its importance, no empirical research has been conducted to determine why the poll failed. Using data from a 1937 Gallup survey which asked about participation in the *Literary Digest* poll I conclude that the magazine's sample and the response were both biased and jointly produced the wildly incorrect estimate of the vote. But, if all of those who were polled had responded, the magazine would have, at least, correctly predicted Roosevelt the winner. The current relevance of these findings is discussed.

Samples were collected through telephone calls. Since telephones were expensive, the sampling method biased wealthier voters

	Landon (Rep)	Roosevelt (Dem)
Predicted	57%	43%
Actual	38%	62%

Quota Sample

 New York Times

Trump Repeats Truman? Not Quite - NYTimes.com

George Gallup kept polling until mid-October, but then rested on those ... In the popular vote, he took 49.6 percent to Dewey's 45.1 percent.

Nov 9, 2016

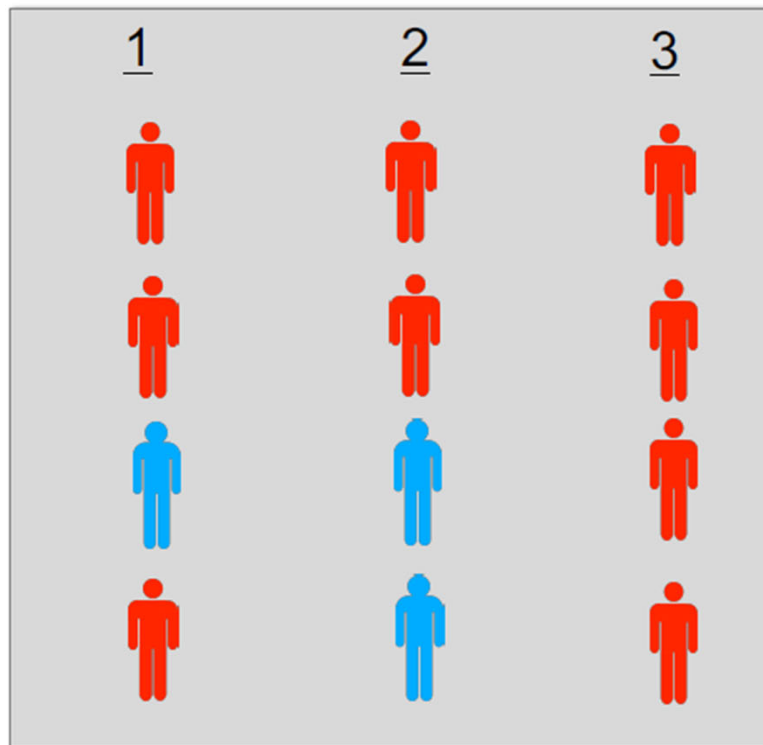


Interviewers were told that they could interview whomever they wished, granted they fulfilled their quotas.

	Dewey (Rep)	Truman (Dem)
Predicted	49.5%	44.5%
Actual	45.1%	49.6%

Bias

Pop.



$$\begin{aligned} P(S1) &= 0.5 \\ P(S1|R) &= 0.75 \\ P(S1|B) &= 0.25 \end{aligned}$$



$$\begin{aligned} P(S2) &= 0.5 \\ P(S2|R) &= 0.5 \\ P(S2|B) &= 0.5 \end{aligned}$$



$$\begin{aligned} P(S3) &= 0.5 \\ P(S3|R) &= 1 \\ P(S3|B) &= 0 \end{aligned}$$



Bias

Pop.



Exhibit 2
Source of Sample: Random Digit Dialing

Category	Frequency
Disconnected/non working number	702
Business/non-household number	237
No answer, busy signal, not at home	1467
Interviewer reject - language barrier	15
Interviewer reject - other	8
Respondent refusal	189
Ineligible respondent	531
Termination by respondent	1
Completed interviews	<u>99</u>
Total	3249

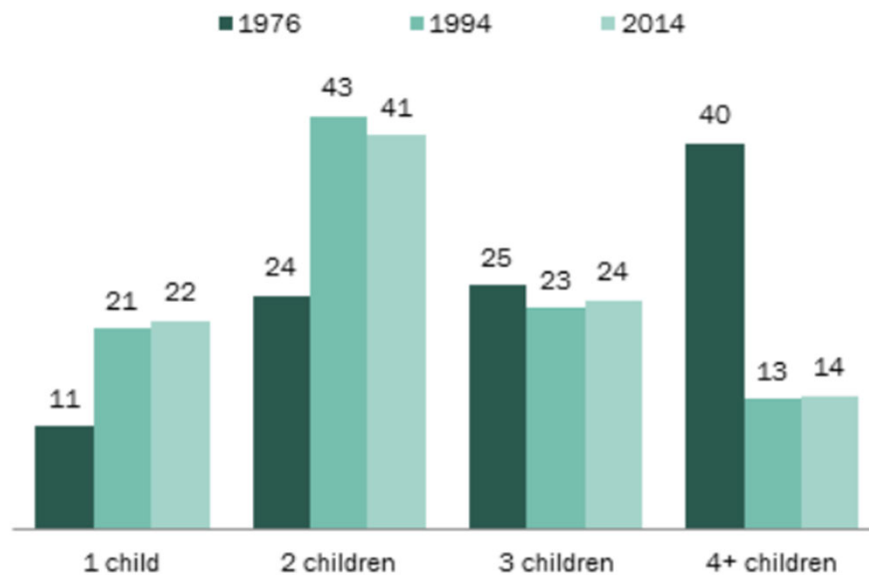


Bias

Pew Research Study
Fertility and Education

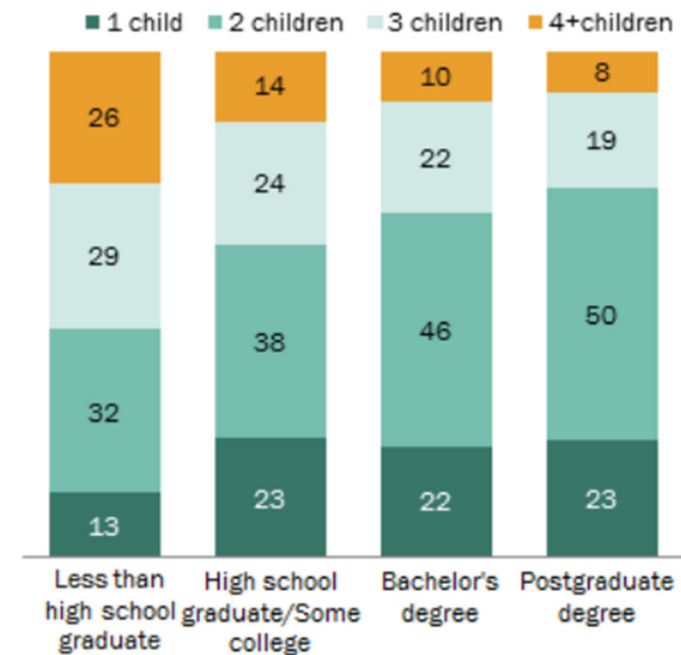
Among Mothers, Family Size is Shrinking

% of mothers ages 40 to 44 with...



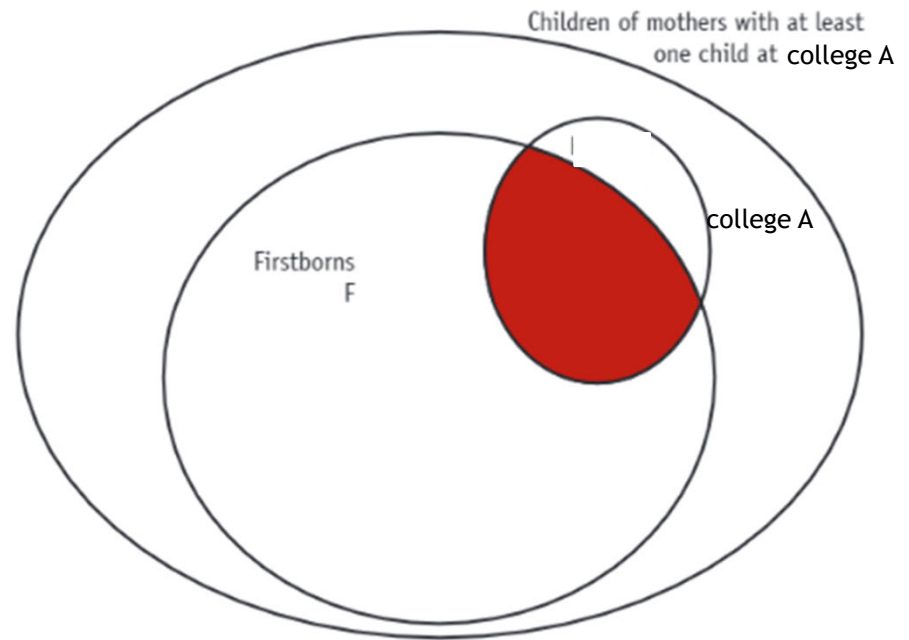
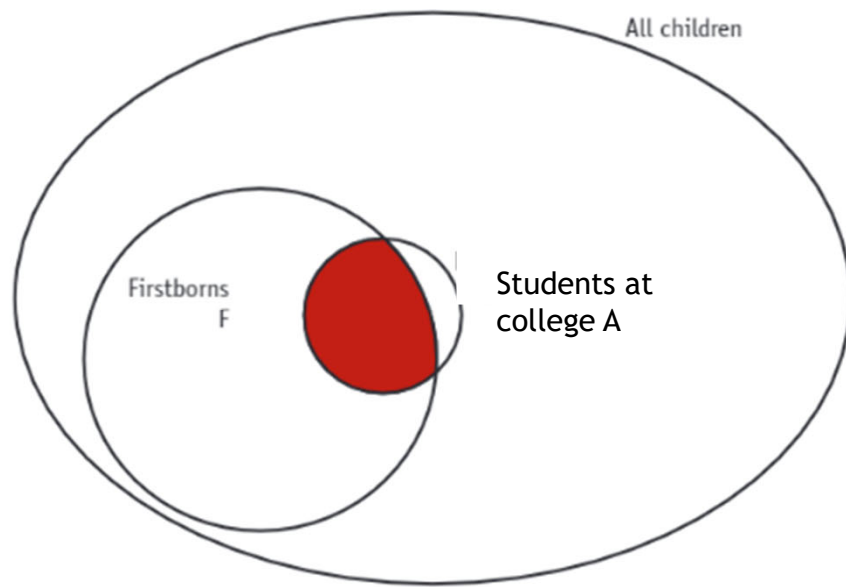
Moms with Less Education Have Bigger Families

% of mothers ages 40 to 44 with ...

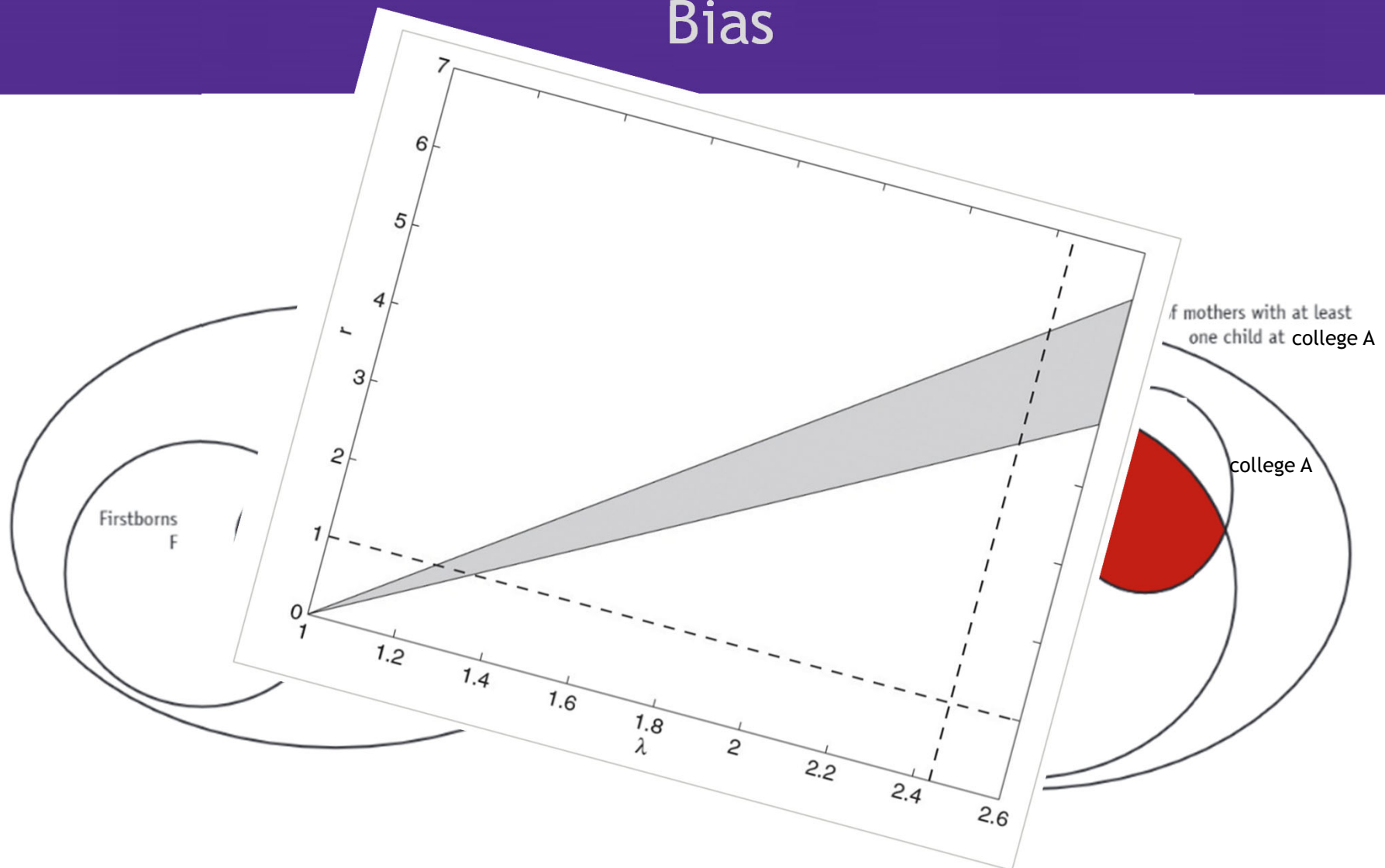


Bias

Base Rate Fallacy



Bias



Questions

- ▶ Questions on Piazza?
- ▶ Question for You!

What are the advantages and drawbacks of automated grading?

Thought Experiment: What Are the Ethical Implications of a Robo-Grader?

Will asked students to consider whether they would want their essays automatically graded by an underlying computer algorithm, and what the ethical implications of automated grading would be. Here are some of their thoughts.

Questions

► Questions on Piazza?

► Question for You!

What are the advantages and drawbacks of automated grading?

TL
In []:

```
def is_prime(x):  
    # YOUR CODE HERE  
    raise NotImplementedError()
```

In []:

```
# TEST  
assert is_prime(2)==True  
assert is_prime(4)==False
```

In []: