Center for Data Science
NYU

DS-UA 112
Introduction to Data Science

Week 3: Lecture 1
Tables – Arranging Data in Rows and Columns

Center for Data Science
NYU

How can tables help us to summarize data?

DS-UA 112
Introduction to Data Science

Week 3: Lecture 1

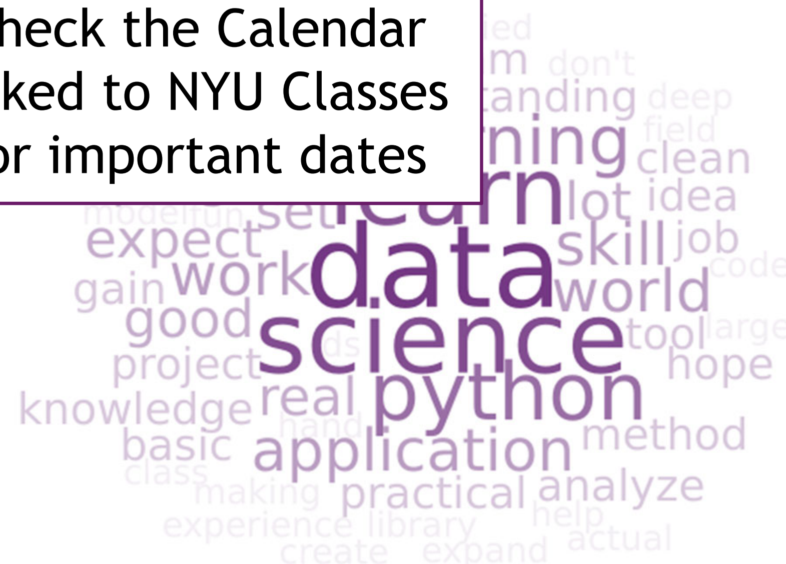
Tables – Arranging Data in Rows and Columns

Adapted from Nolan, Hug, and Salganik

Announcements

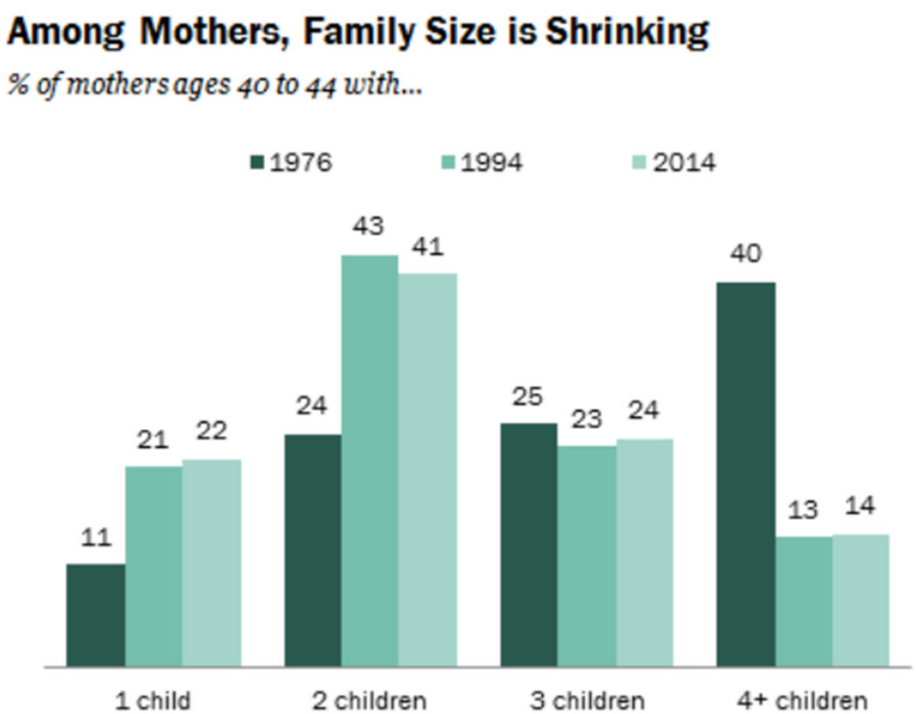
- Please check Week 3 agenda on NYU Classes
  - Homework 1
  - Lab 3
  - Grader Office Hours
- Remember to post to Piazza

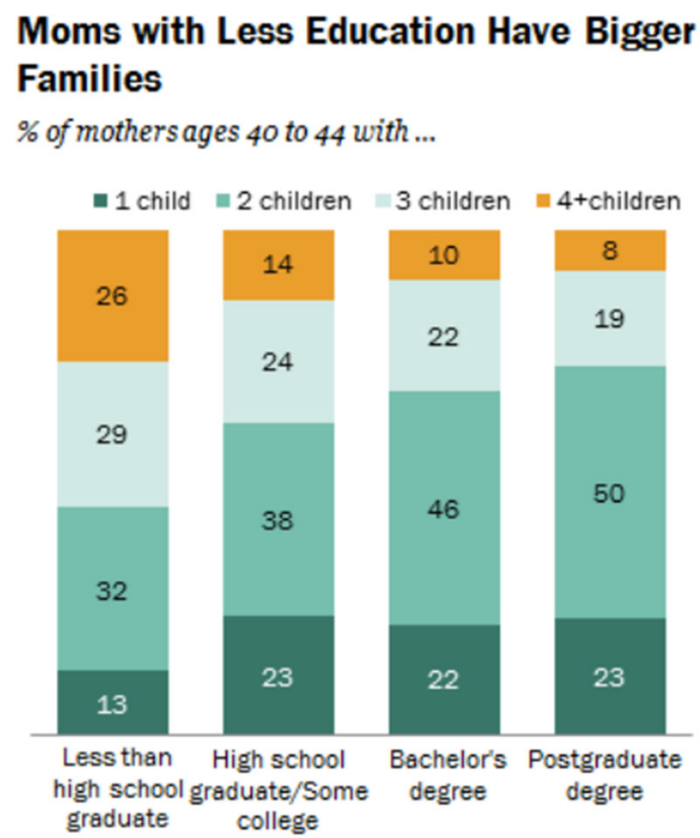
Check the Calendar linked to NYU Classes for important dates



Review

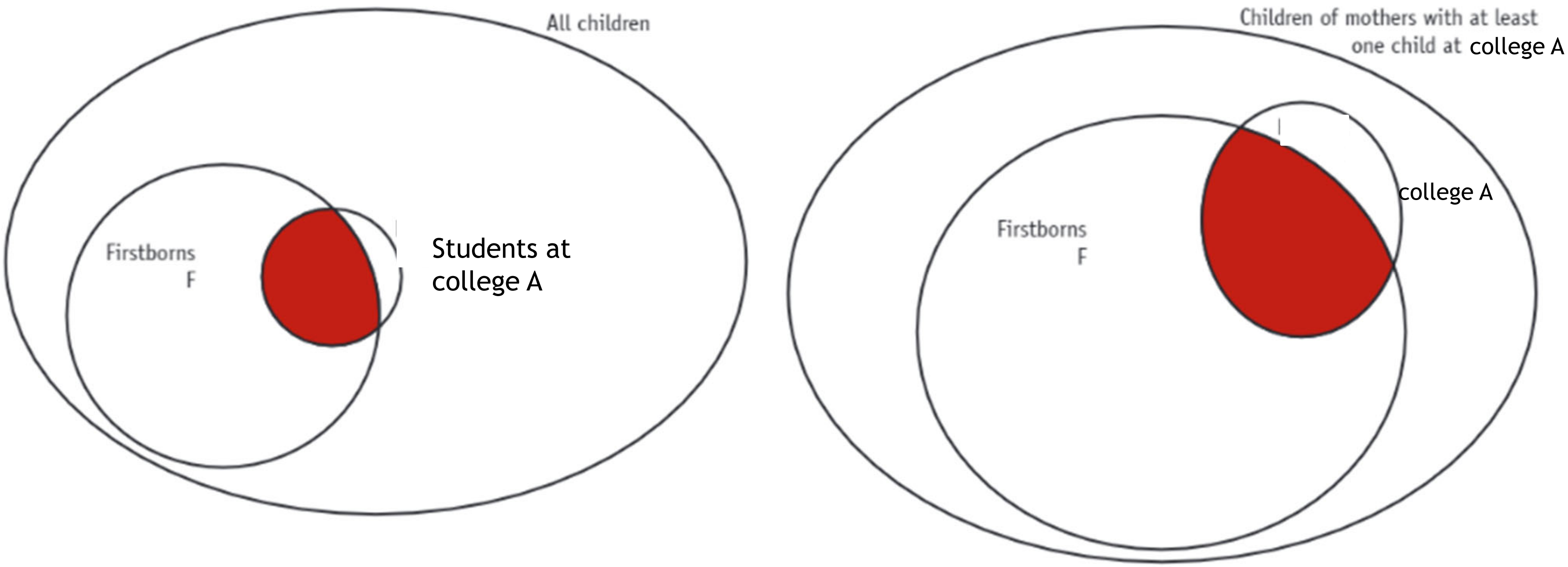
Pew Research Study
Fertility and Education



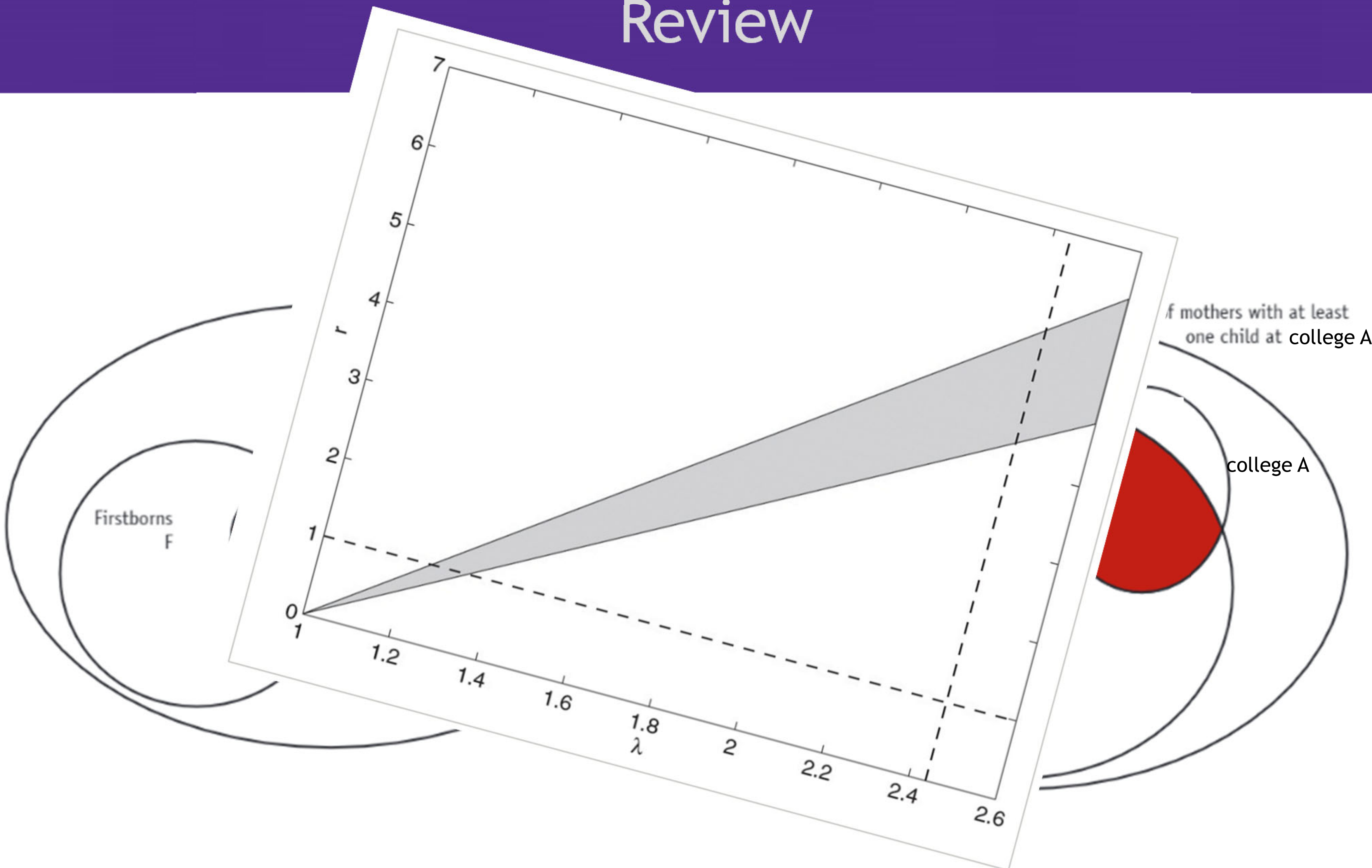


Review

Base Rate Fallacy



Review



Firstborns
F
f mothers with at least
one child at college A
college A
7
6
5
4
3
2
1
0
r
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
λ

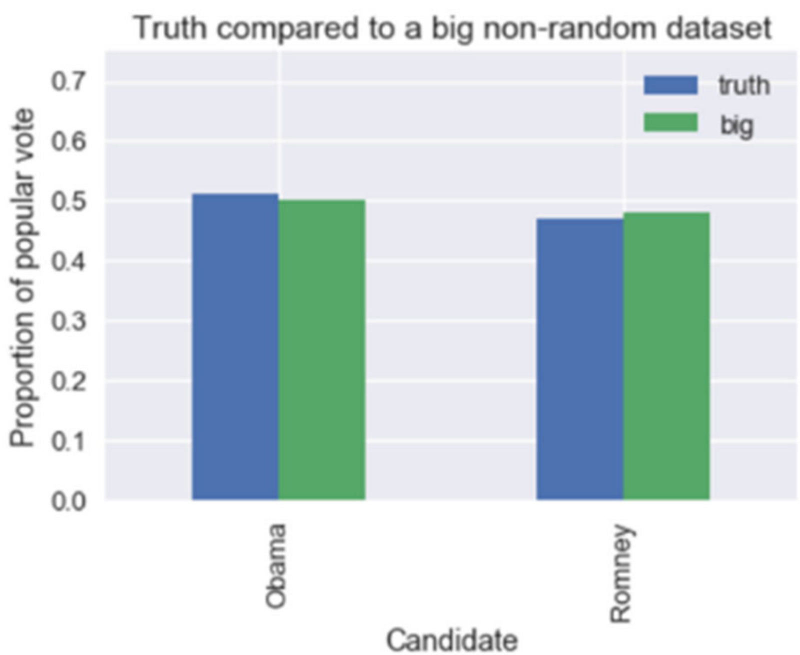
Agenda

- Probability
  - Addition, Multiplication, Complement Rules
  - Summarize with average value
- Confounded Data
  - Adjusting for Bias
- Messy Data
  - Arranging into Rows and Columns

References

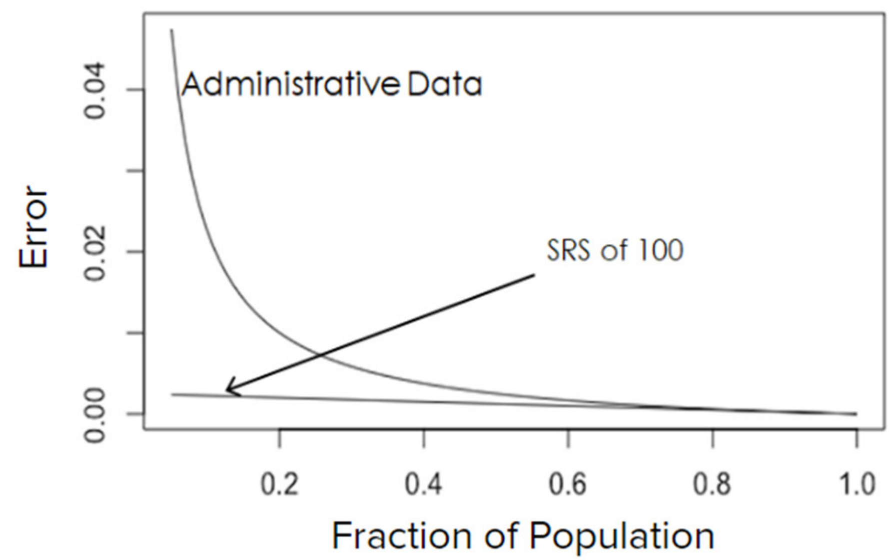
- Nolan, Lau, Gonzalez (Chapter 2, 3.1)
  - https://cp71.github.io/textbook
- Salganik (Chapter 3)

Not Representative



If a dataset is not representative, then it may or may not be suitable for a study. Sometimes it causes bias in the analysis.

Approach to data collection could indicate bias to us.



Not Representative

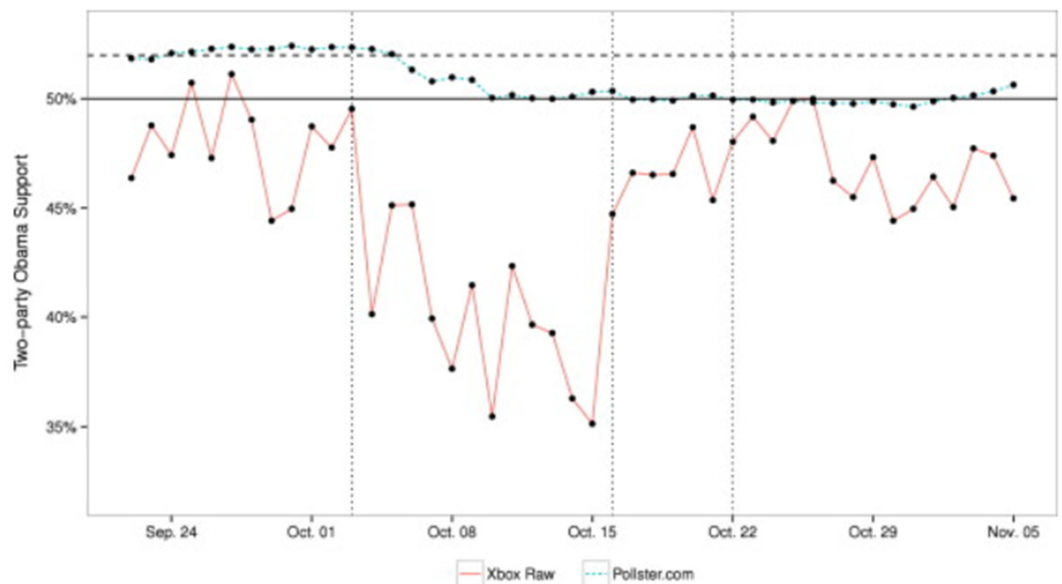
Sampling Frame may not lie in the Population

Forecasting elections with non-representative polls

Wei Wang [a], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a, c]

Abstract

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After



Not Representative

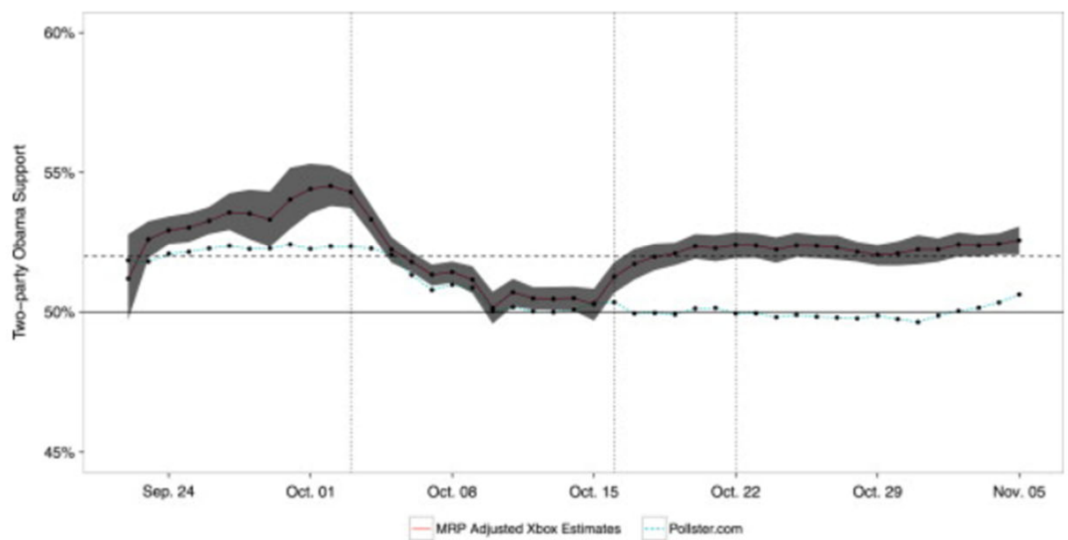
Adjustments made by stratifying following the data collection

Forecasting elections with non-representative polls

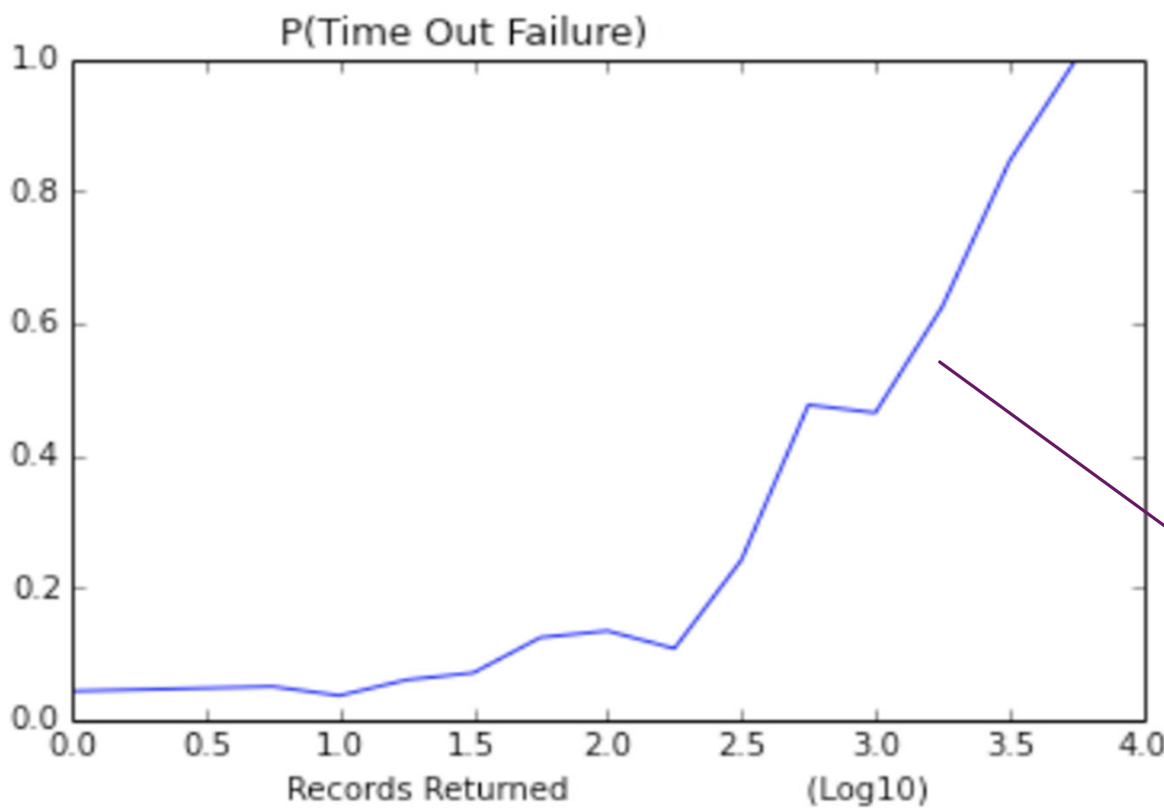
Wei Wang [a], David Rothschild [b], Sharad Goel [b], Andrew Gelman [a, c]

Abstract

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After



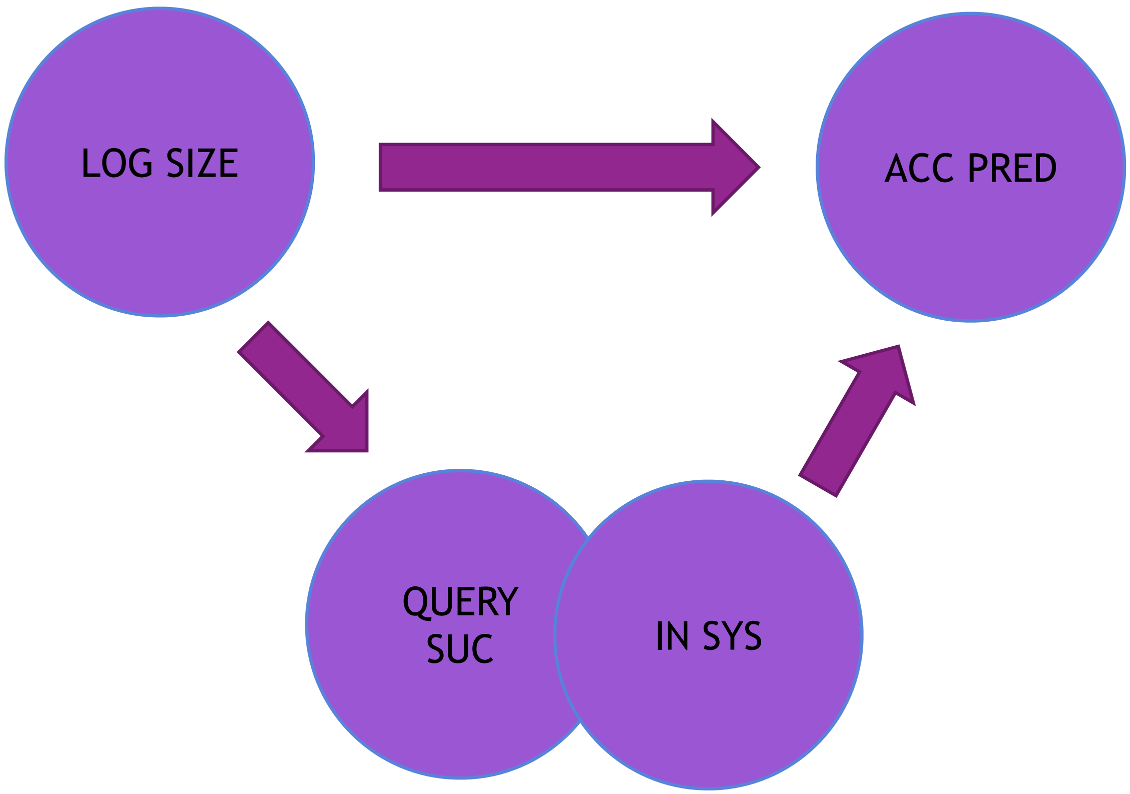
Confounded Data



Here the probability of inclusion of the record in the sample changes depending on the size of the record

How could this lead to bias? In particular, why might predictions reflect habits of new customers

Confounded Data



ACC PRED
    Accurate Prediction of Least Popular
LOG SIZE
    Size of Database Record
QUERY SUC
    Whether Query Returned Successfully
IN SYS
    Whether Record in the System

Confounded Data

How to adjust the probabilities?

P(ACC PRED | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC) x P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE,NOT QUERY SUC) x P(NOT QUERY SUC | LOG SIZE) =

Confounded Data

How to adjust the probabilities?

P(ACC PRED | LOG SIZE) =

P(ACC PRED | LOG SIZE, QUERY SUC) x P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE, NOT IN SYS) x P(NOT QUERY SUC | LOG SIZE) =

Confounded Data

How to adjust the probabilities?

P(ACC PRED | LOG SIZE)

=

P(ACC PRED | LOG SIZE, QUERY SUC) x P(QUERY SUC | LOG SIZE)

+

P(ACC PRED | LOG SIZE, NOT IN SYS) x P(NOT QUERY SUC | LOG SIZE)

=

P(ACC PRED | LOG SIZE, QUERY SUC)P(QUERY SUC | LOG SIZE)

+

(0) P(NOT QUERY SUC | LOG SIZE)

Confounded Data

How to adjust the probabilities?

P(ACC PRED | LOG SIZE)

=

P(ACC PRED | LOG SIZE, QUERY SUC) x P(QUERY SUC | LOG SIZE)

Without this quantity the two sides would not be equal.

Confounded Data

How to adjust the probabilities?

P(ACC PRED | LOG SIZE)

=

P(ACC PRED | LOG SIZE, QUERY SUC) x P(QUERY SUC | LOG SIZE)

Without this quantity the two sides would not be equal.

$$\frac{\text{P(ACC PRED | LOG SIZE)}}{\text{P(QUERY SUC | LOG SIZE)}}$$

=

P(ACC PRED | LOG SIZE, QUERY SUC)

Confounded Data

Gender Achievement Gaps in U.S. School Districts

Author/s: Sean F. Reardon , Erin Fahle , Demetra Kalogrides , Anne Podolsky , Rosalía C. Zárate

Year of Publication: 2018

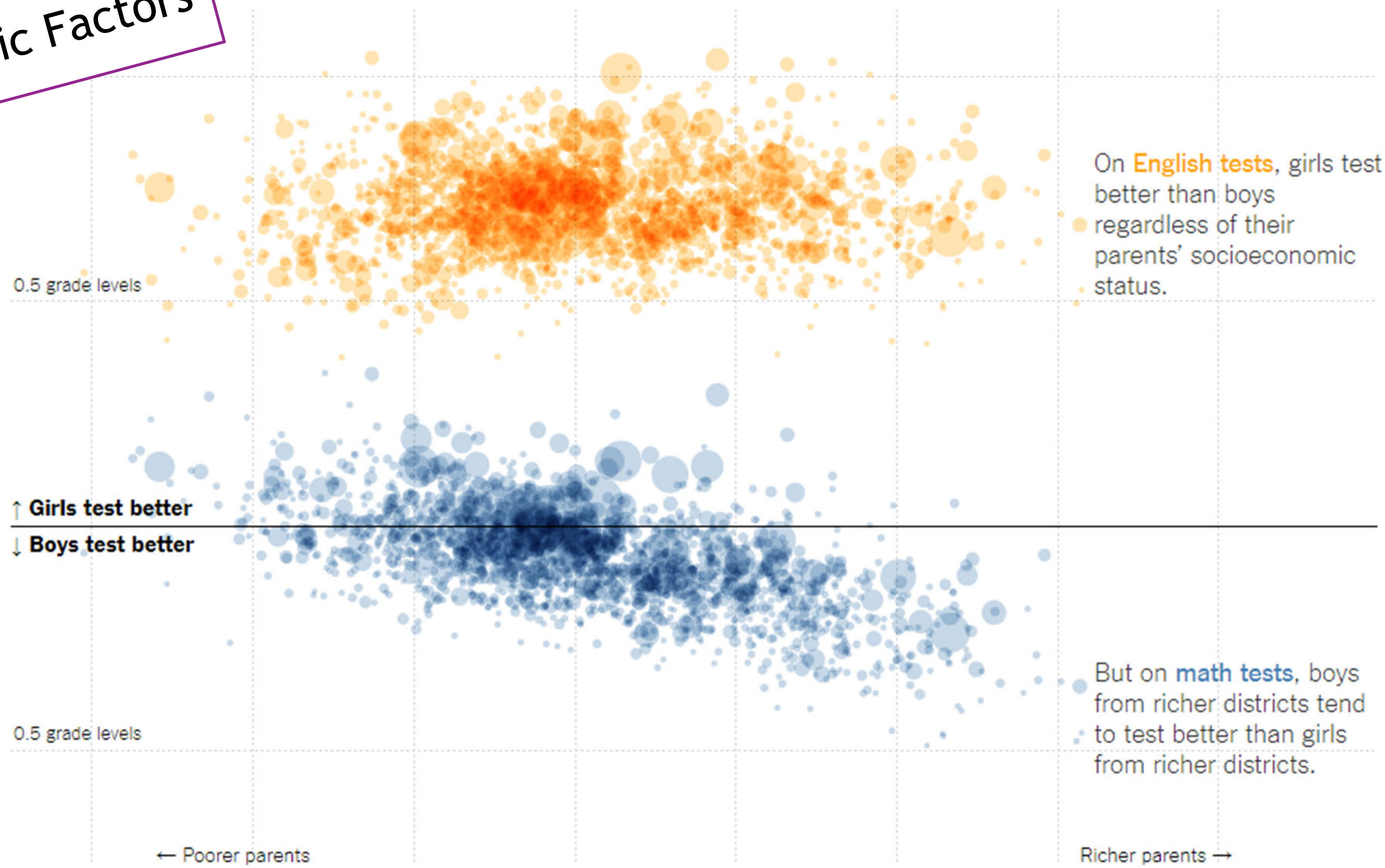
In the first systematic study of gender achievement gaps in U.S. school districts, we estimate male-female test score gaps in math and English Language Arts (ELA) for nearly 10,000 school districts in the U.S. We use state

- What is the population?
- What is the question under study?

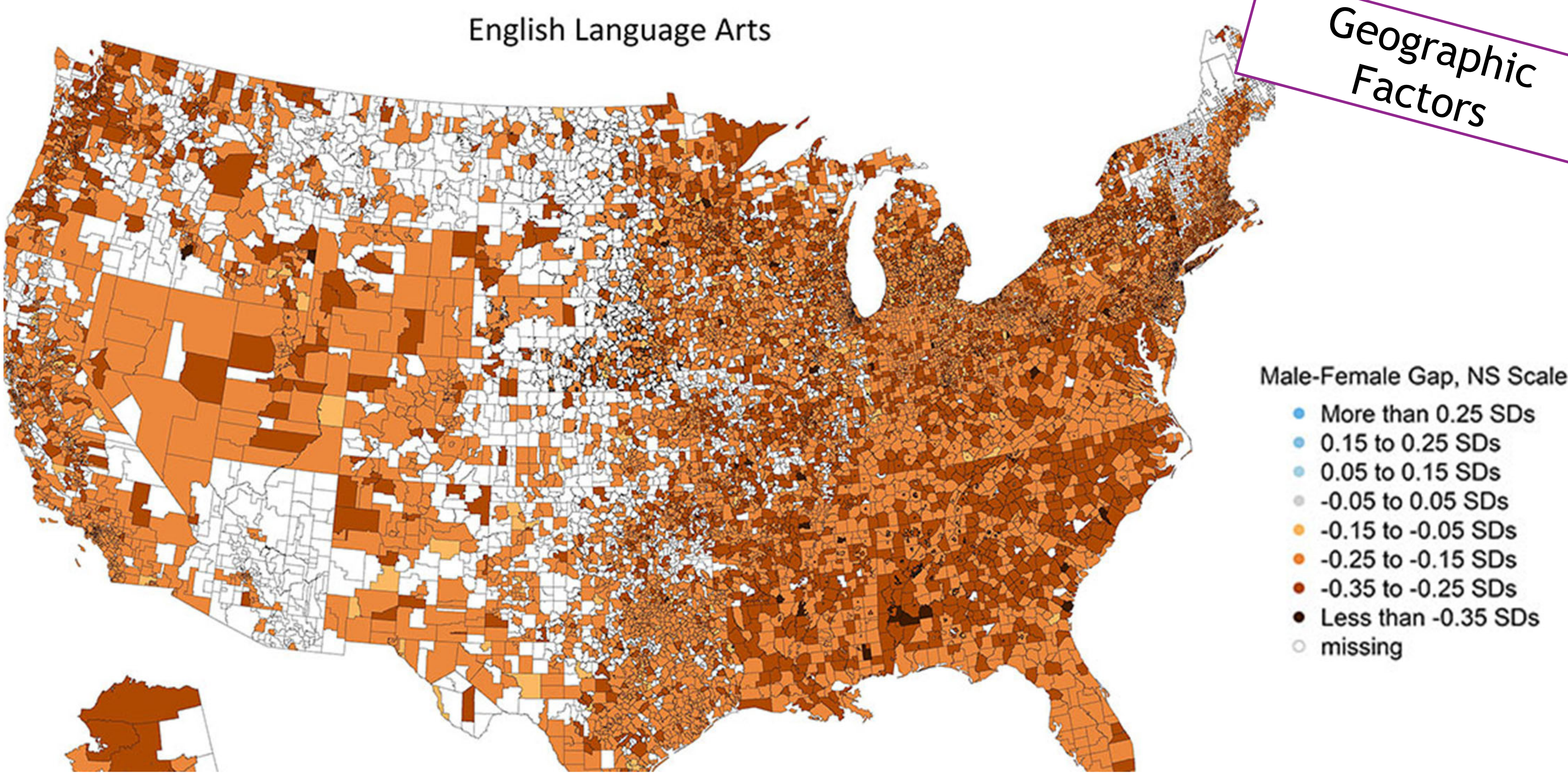
- What is the sampling frame?
- What could lead to confounded data?

Confounded Data

Economic Factors



Confounded Data



Geographic Factors

Questions

- Questions on Piazza?
- Question for You!

Should the word data be understood as singular or plural?

In Latin, data is the plural of datum and, historically and in specialized scientific fields , it is also treated as a plural in English, taking a plural verb, as in the data were collected and classified . In modern non-scientific use, however , despite the complaints of traditionalists, it is often not treated as a plural. Instead, it is treated as a mass noun, similar to a word like information, which cannot normally have a plural and which takes a singular verb. Sentences such as data was (as well as data were) collected over a number of years are now widely accepted in standard English.

Questions

- Questions on Piazza?
- Question for You!

Should the word data be understood as singular or plural?