Data Science 100

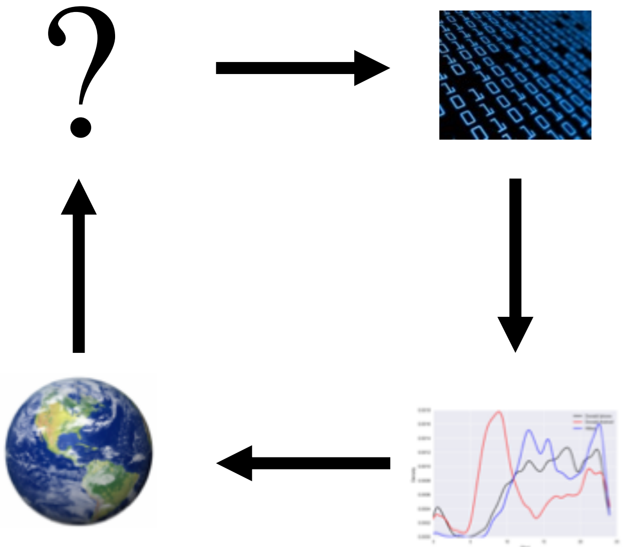
Principles & Techniques of Data Science

Slides by:

Deborah Nolan deborah_nolan@berkeley.edu

Spring 2018 updates

Fernando Pérez fernando.perez@berkeley.edu



Announcements for Today

- Midterm is scheduled for **March 8 during class**
- We will try using Google forms today, but may need to resort to clickers/cell phone
- Slides and notes from lecture available online at http://ds100.org/sp18
- HW 1 will be released next Tuesday (Jan 23).

A quick bit about me

- New faculty in statistics department (Fall 2017)
- Physics, applied math background
- Open source tools for science, reproducible research, interactive data science

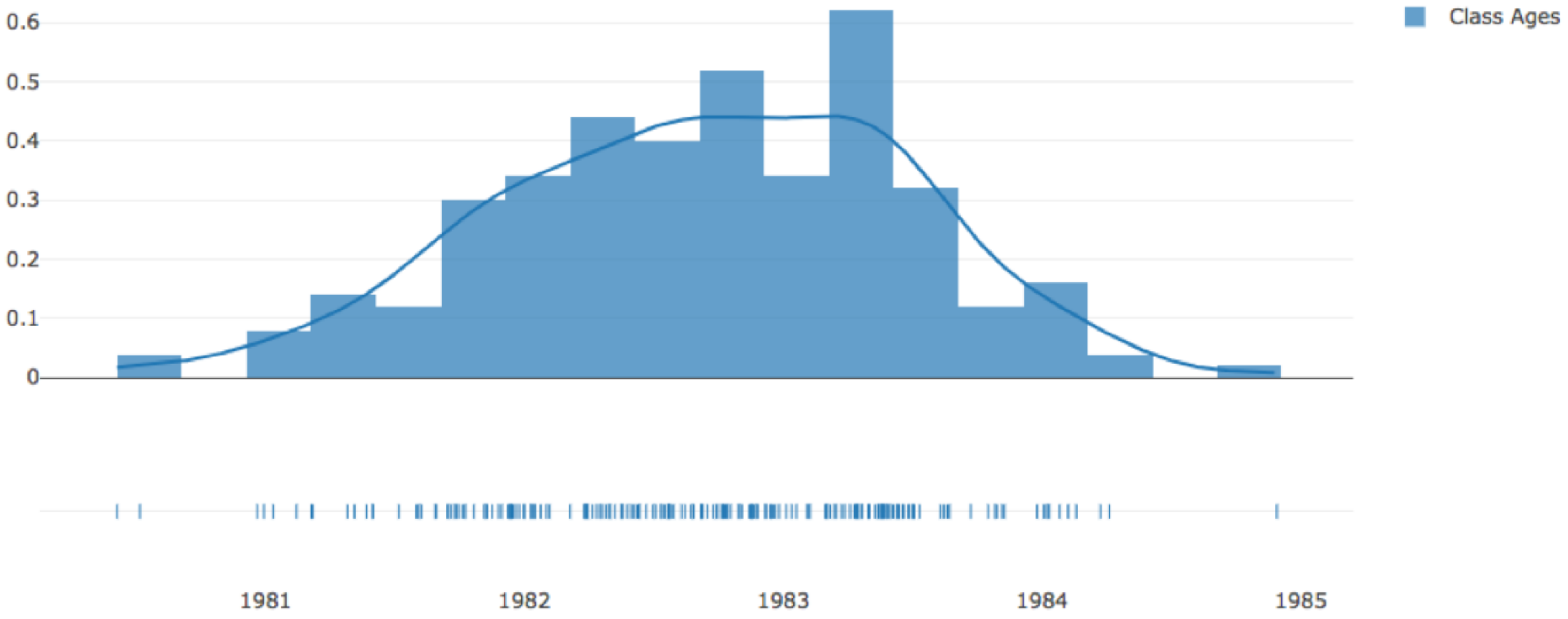
Fernando Pérez
419 Evans
OH: Tues 2-4pm

IP[y]:
IPython

jupyter

Over 1,400 contributors!

Let’s finish name/age model from Tues.



Three themes in DS 100

- *Problem formulation*
- Data Provenance
- Implications of scale

Each class will address one or more of these themes

Data Provenance

We take a narrow approach to this topic

How were these data generated?

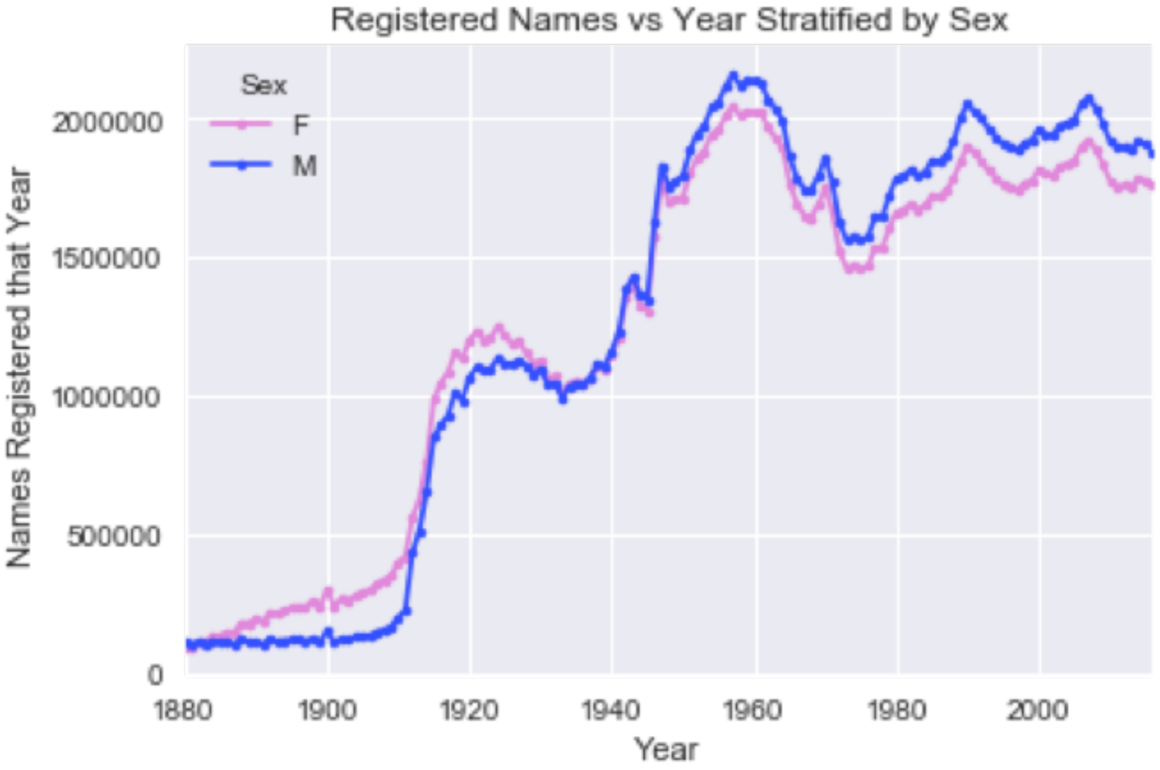
- The mechanisms by which the data arose impacts whether we can answer the question of interest
- Can we generalize beyond what we observe in our data?

How were these data generated?

- Do they form a census – information on all subjects?
- Are they a subset or sample?
  - Was a chance mechanism used to select the sample?
  - Or, is this a self-selected, convenience, or judgment sample?
  - Or, an administrative database?
- Were the data generated from
  - a collection in nature
  - an industrial process
  - a social science study

Baby Names Data

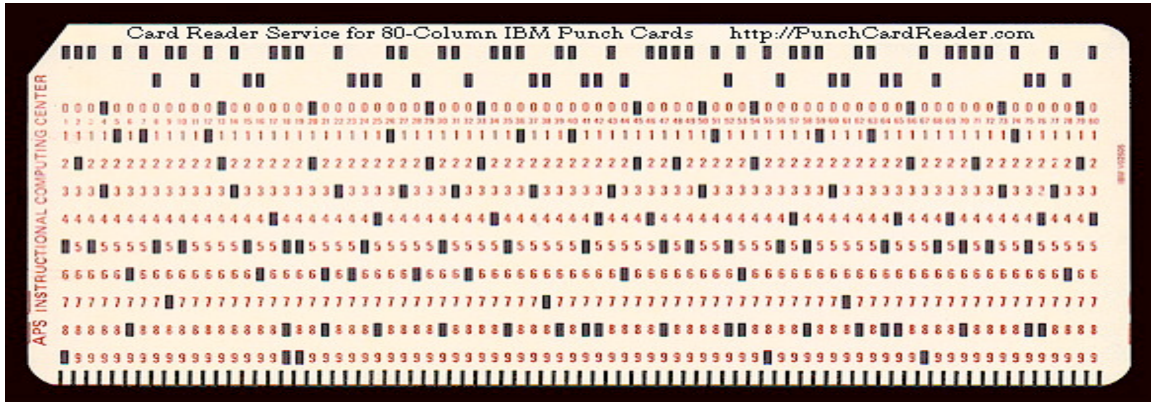
- Do they form a census?
- Can we use our findings to make generalizations about the students in this class?
- SSN system created 1937

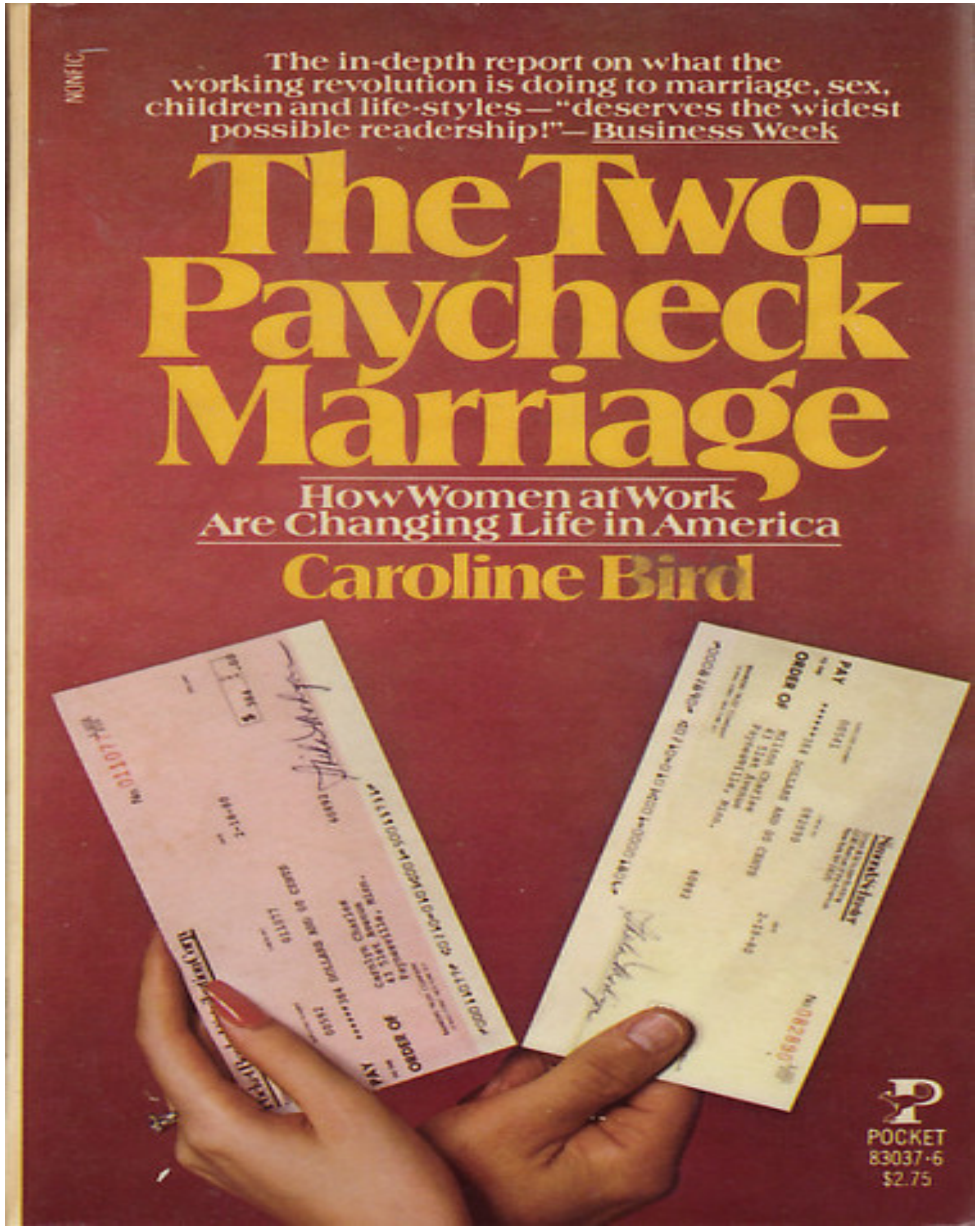


Statistical Science – A first encounter (D. Nolan)

Data collected from questionnaire published in Women's Day magazine

Thousands of women responded to questions about their lifestyle


Card Reader Service for 80-Column IBM Punch Cards
http://PunchCardReader.com
APS INSTRUCTIONAL COMPUTING CENTER


NONFIC
The in-depth report on what the working revolution is doing to marriage, sex, children and life-styles—"deserves the widest possible readership!"—Business Week
The Two-Paycheck Marriage
How Women at Work Are Changing Life in America
Caroline Bird
POCKET
83037-6
$2.75

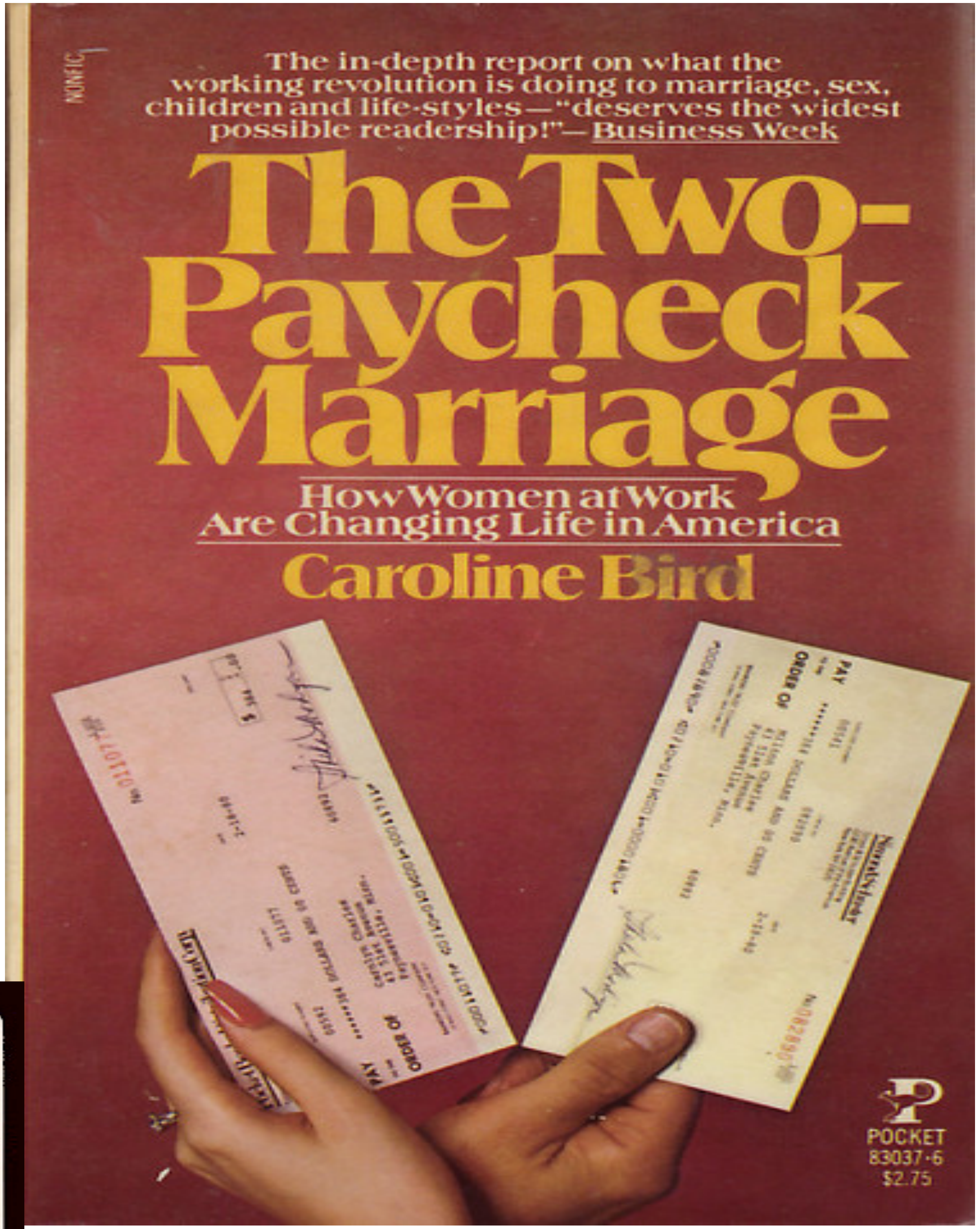
Statistical Science – A first encounter

Do they form a census – information on all subjects?

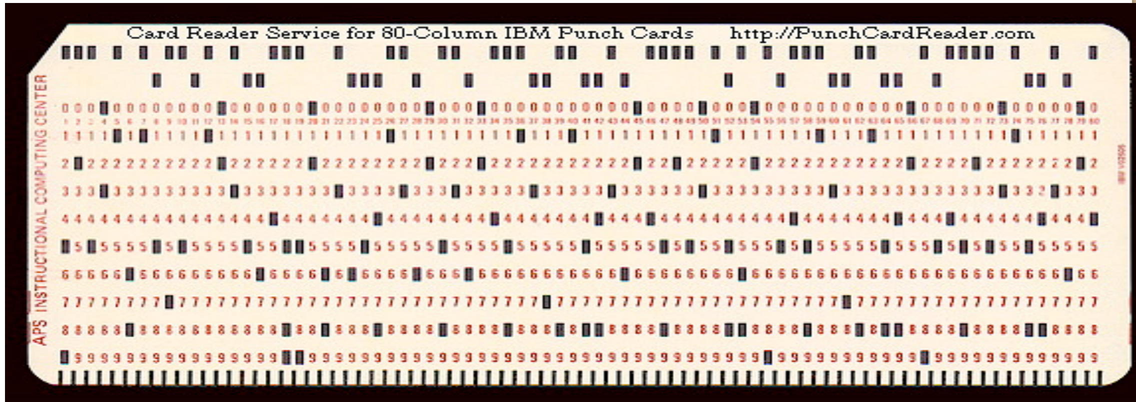
NOPE

Were there enough responses to overcome any issues with a nonprobability sample?

NOPE



NONFIC
The in-depth report on what the working revolution is doing to marriage, sex, children and life-styles—"deserves the widest possible readership!"—Business Week
The Two-Paycheck Marriage
How Women at Work Are Changing Life in America
Caroline Bird
POCKET
83037-6
$2.75



Card Reader Service for 80-Column IBM Punch Cards
http://PunchCardReader.com
APS INSTRUCTIONAL COMPUTING CENTER

Computational Science: powerful tools

TensorFlow
?
jupyter
APACHE Spark™
scikit learn
julia
R
python™
sas®

The purpose of computing is insight, not numbers.

R. Hamming
Numerical Methods for Scientists and Engineers (1962)

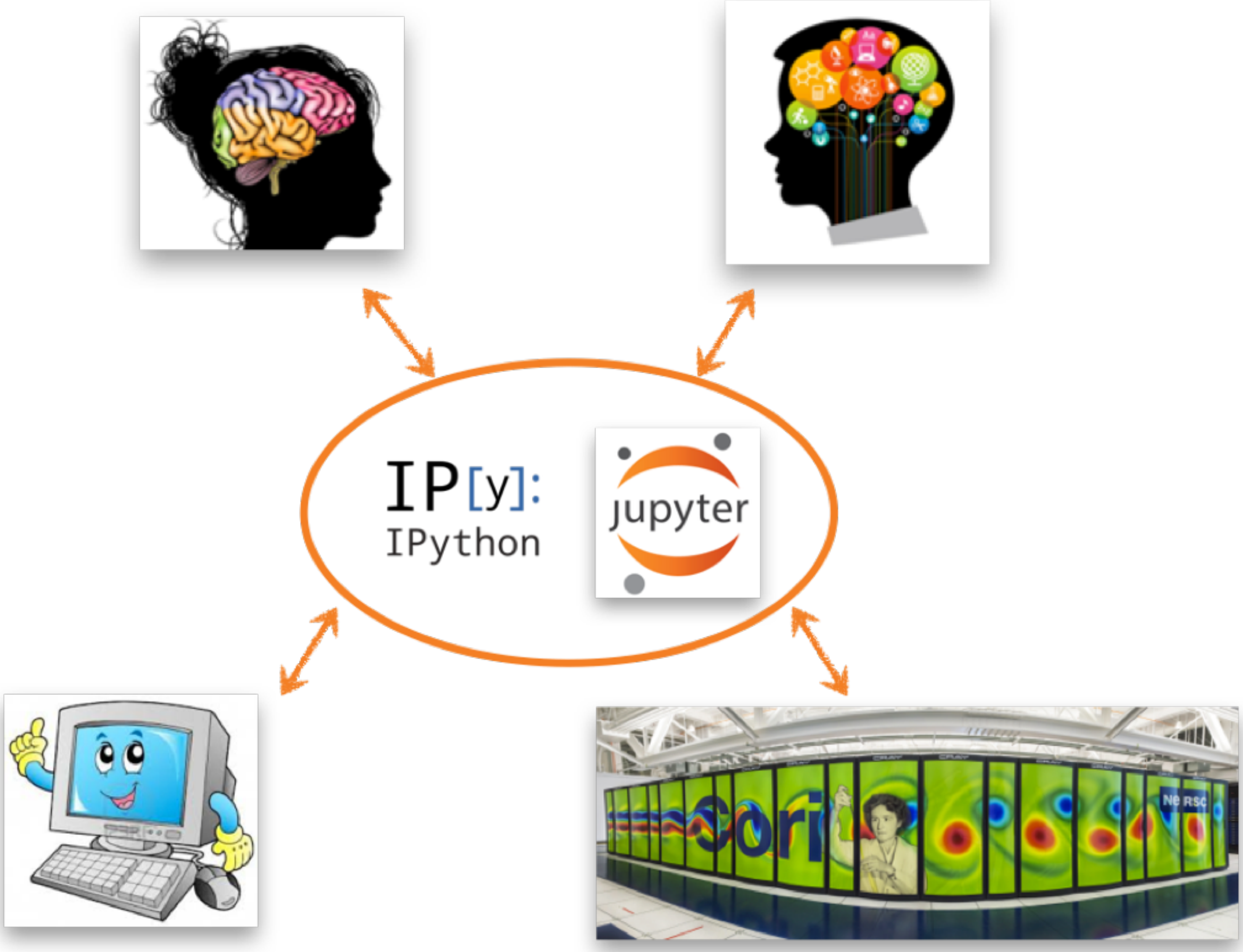
Lessons Learned

Good Data Analysis ≠

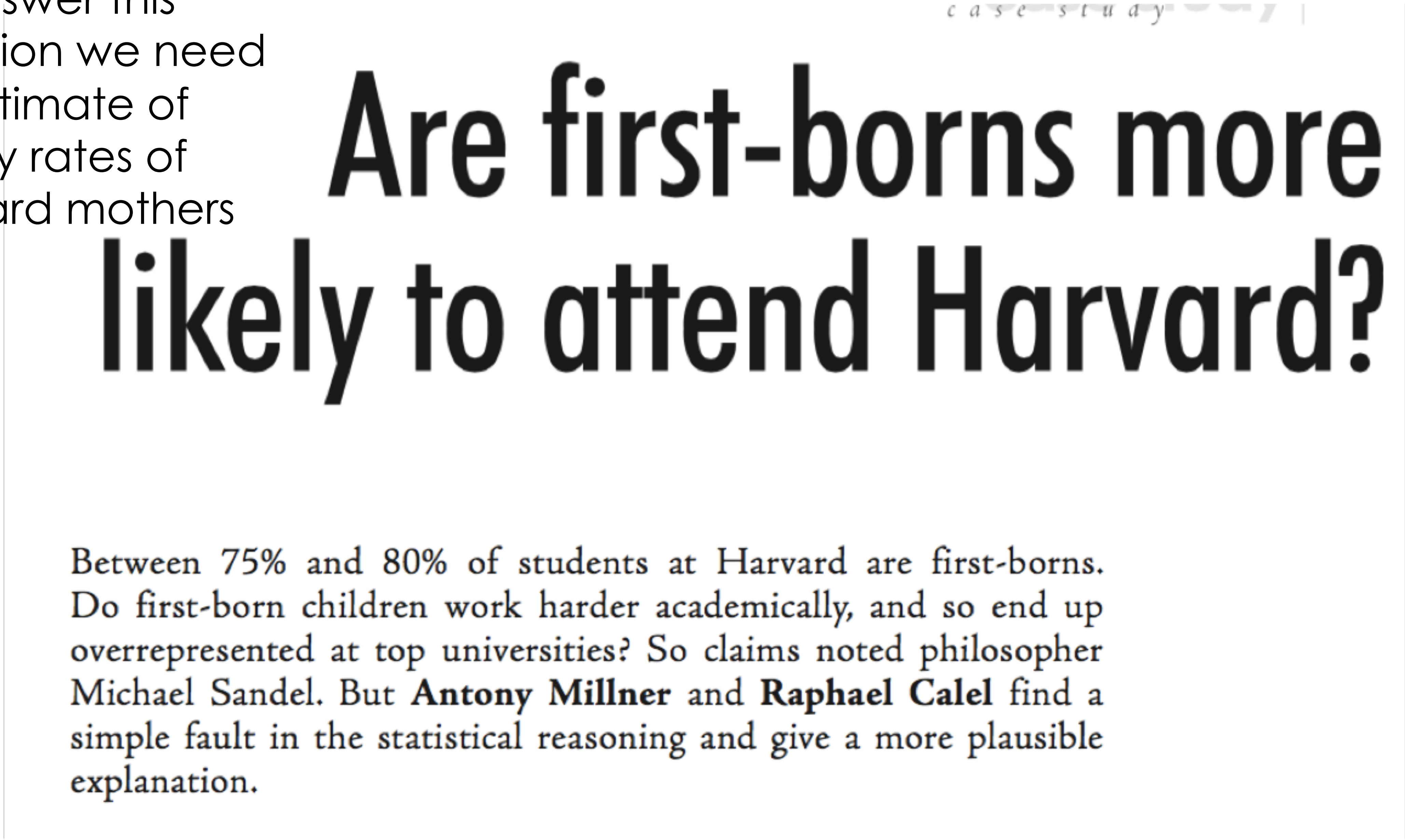
Simple Application of a Statistics Recipe

Good Data Analysis ≠

Simple Application of Statistical Software


IP[y]:
IPython
jupyter
Cori
NERSC

To Answer this question we need an estimate of fertility rates of Harvard mothers



case study

Are first-borns more likely to attend Harvard?

Between 75% and 80% of students at Harvard are first-borns. Do first-born children work harder academically, and so end up overrepresented at top universities? So claims noted philosopher Michael Sandel. But **Antony Millner** and **Raphael Calel** find a simple fault in the statistical reasoning and give a more plausible explanation.

DETOUR:

1. What is a simple random sample?
2. Why is it so desirable?
3. Can we make up for not having a SRS with big data?

What is a Simple Random Sample?

Wha

The Simple Random Sample

- *Suppose we have a population with N subjects*
- *We want to sample **n** of them*
- *The SRS is a random sample where every possible unique subset of **n** subjects has the same chance of appearing in the sample*
- This means each person is equally likely to be in the sample

Why is the SRS so Useful?

The Advantages of a SRS

- *Representative: The sample tends to look like the population*

- *Statistics based on the sample tend to be close to statistics based on the population*

- *We can provide typical deviations of sample statistics from population values.*

- *AND MORE…*

Fertility of mothers aged 40-44

Fertility of mothers aged 40-44

- *8.7 million mothers aged 40-44 in the US (CPS)*
- *The have 1, 2, 3, or 4+ children (we treat 4+ as 4)*

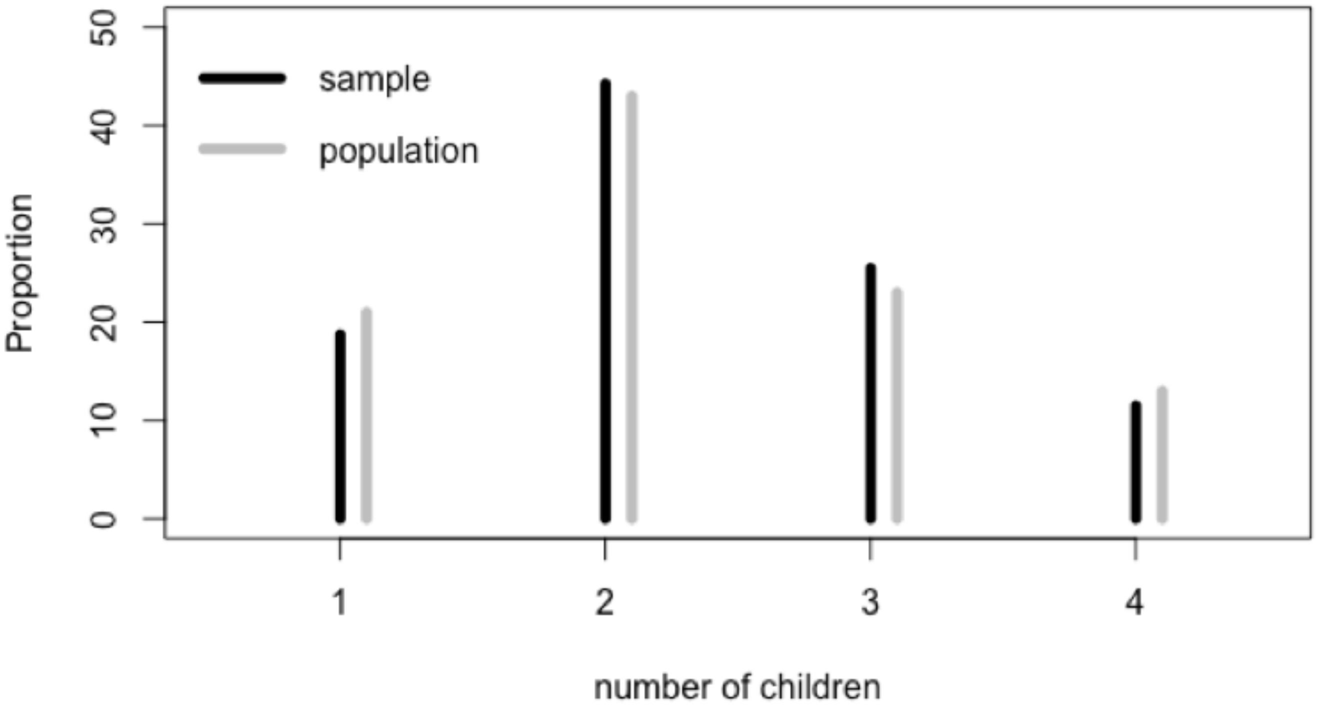
	Number of Children			
	1	2	3	4+
Percent mothers	21%	43%	23%	13%

- *Take a Simple Random Sample of 400 (simulate)*

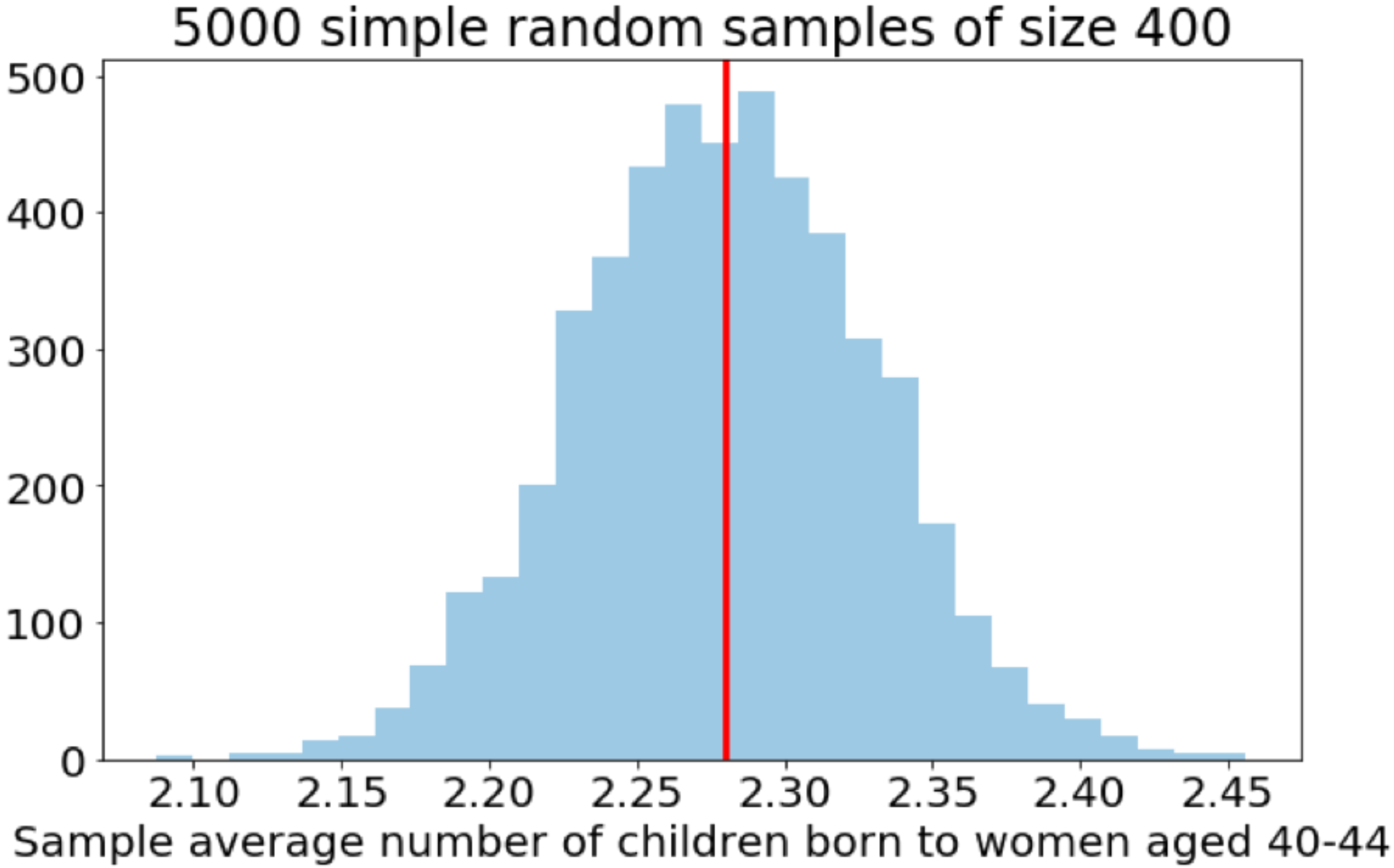
One sample of 400 from 8.7m

Sample AVG = 2.29

Population AVG = 2.28



Repeat



A sample AVG from a SRS will fall close to the true population mean

Can we make up for no SRS with Big Data?

Harvard Professor Meng asked a room full of statisticians at an international conference something like this:

> Suppose we want to estimate a proportion of people in the US that have a particular characteristic. We can either take a SRS of 100 people, or examine an administrative dataset where those who have the characteristic are slightly ore likely to be included in the dataset. How large does the AD need to be to provide a more accurate estimate than the SRS?

***Survey:* [*http://bit.ly/d100-sp18-afd*](http://bit.ly/d100-sp18-afd)**

A. 5% (16m)

B. 20% (64m)

C. 50% (160m)

D. 90% (206m)

More: Meng 2013, *A trio of inference problems that could win you a Nobel Prize in statistics…*

Please also answer the following questions:

- ➢ Are you your birth mother’s first born?

- ➢ How many children does your birth mother have? (including you)

Three possibilities

- SRS of 400 mothers
- SRS of 400 of the children
- An administrative dataset with 80,000 mothers, where those with more children are slightly more likely to be included in the database

Which of these three approaches better?

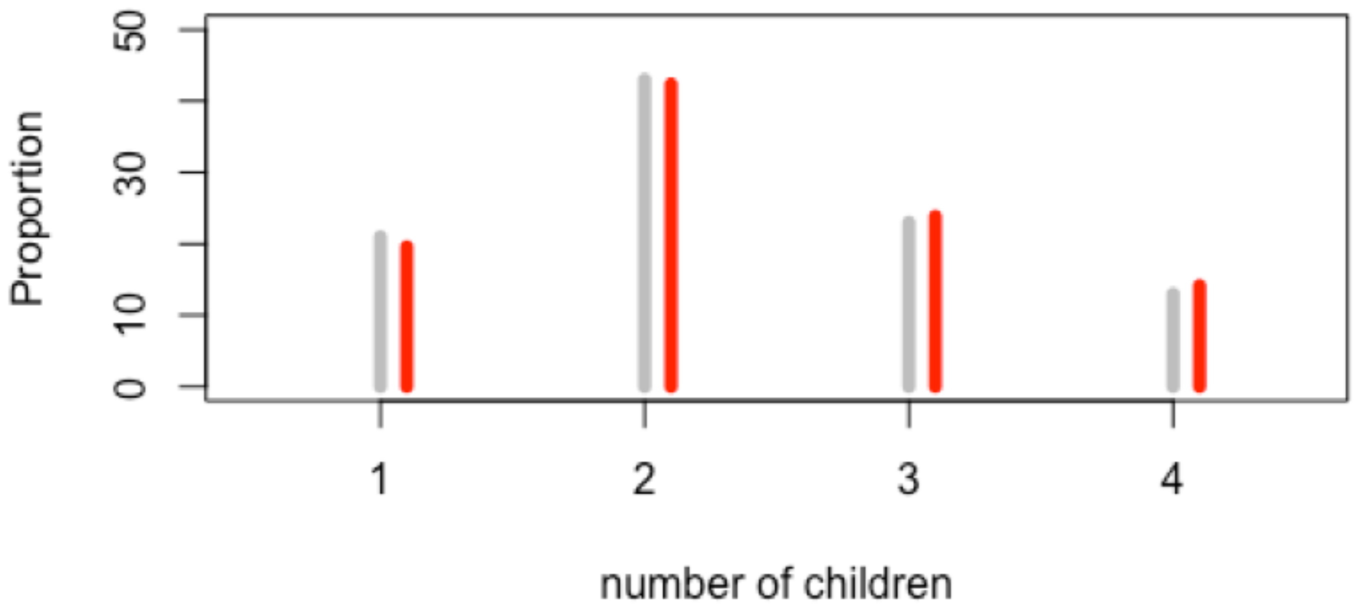
SRS of 400 children

- Compare the chance a child from a 2 children family is chosen to the chance an only child is selected
- What does this imply about the estimate for the average number of children?
- Can we fix this problem?

This is size-biased sampling
Why?
We can fix the problem if we know how many siblings each child has

Administrative Data of 80000 mothers

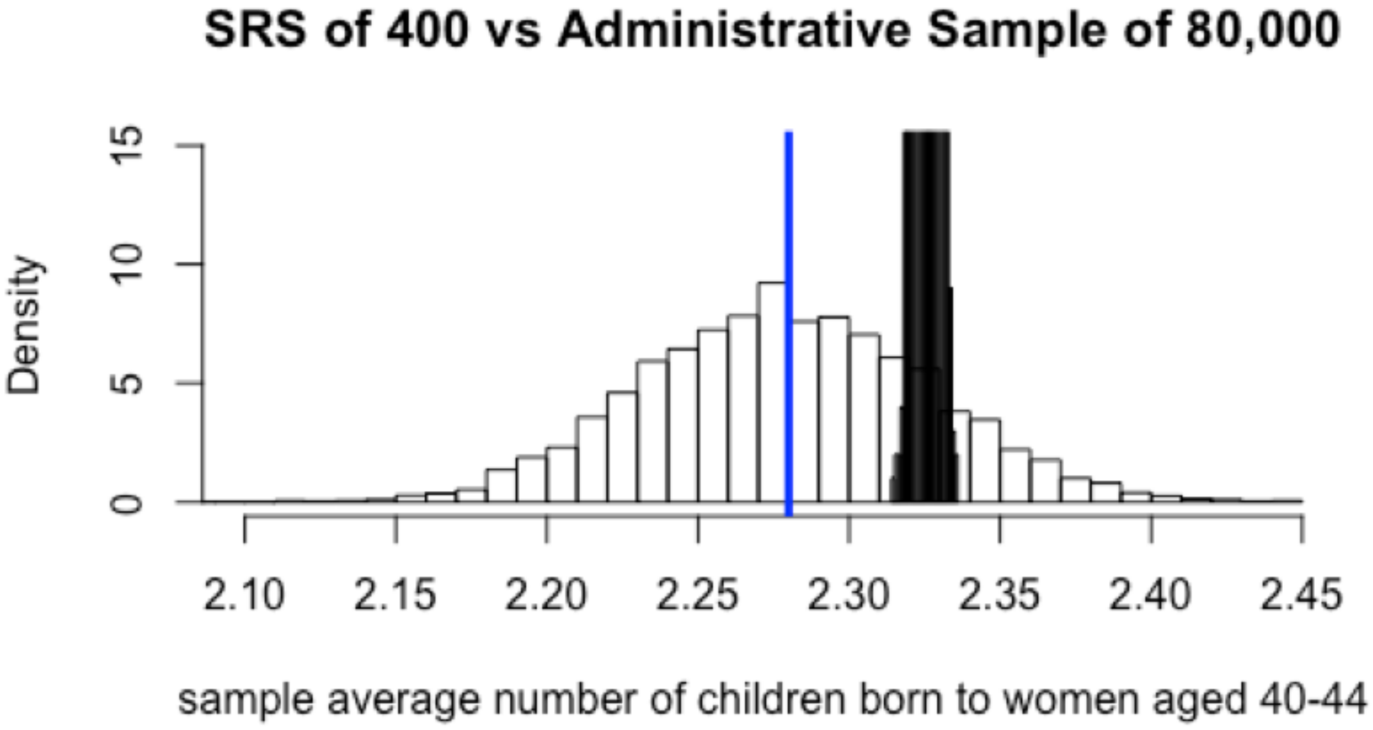
- Compare the proportions in the true population to our administrative dataset



This can't possibly make a difference, right?

Administrative Data of 80,000 mothers

- Compare the accuracy based on 1000 simulations of Administrative Correlated Data



Are first-borns more likely to attend Harvard?

And What about Berkeley?

First-borns predominate at Harvard

- Professor Sandel found 75% of the students in his class are first born
- We know from our study of mothers aged 40-44

	Number of Children			
	1	2	3	4+
Percent mothers	21%	43%	23%	13%

- The proportion of first born children in this group of children: 100/(21 + 86 + 69 + 52) = **1/2.3** = 0.44 or 44%

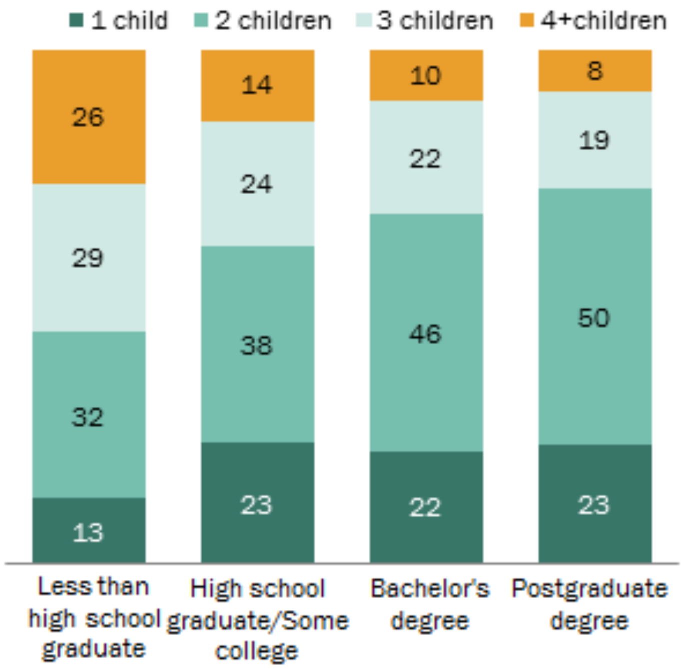
First-borns predominate at Harvard

- BUT that’s the wrong comparison
- What about Data Provenance?
- How large are Harvard families?
- Do we have any reason to believe that mothers of Harvard students might not have the same fertility rate as all US mothers aged 40-44?

According to Pew Research Center

Moms with Less Education Have Bigger Families

% of mothers ages 40 to 44 with ...



Note: High school graduate/Some college includes those with a two-year degree. Postgraduate degree includes those with at least a master's degree. Figures may not add to 100% due to rounding.

Source: Pew Research Center analysis of 2012 and 2014 Current Population Survey June Supplements

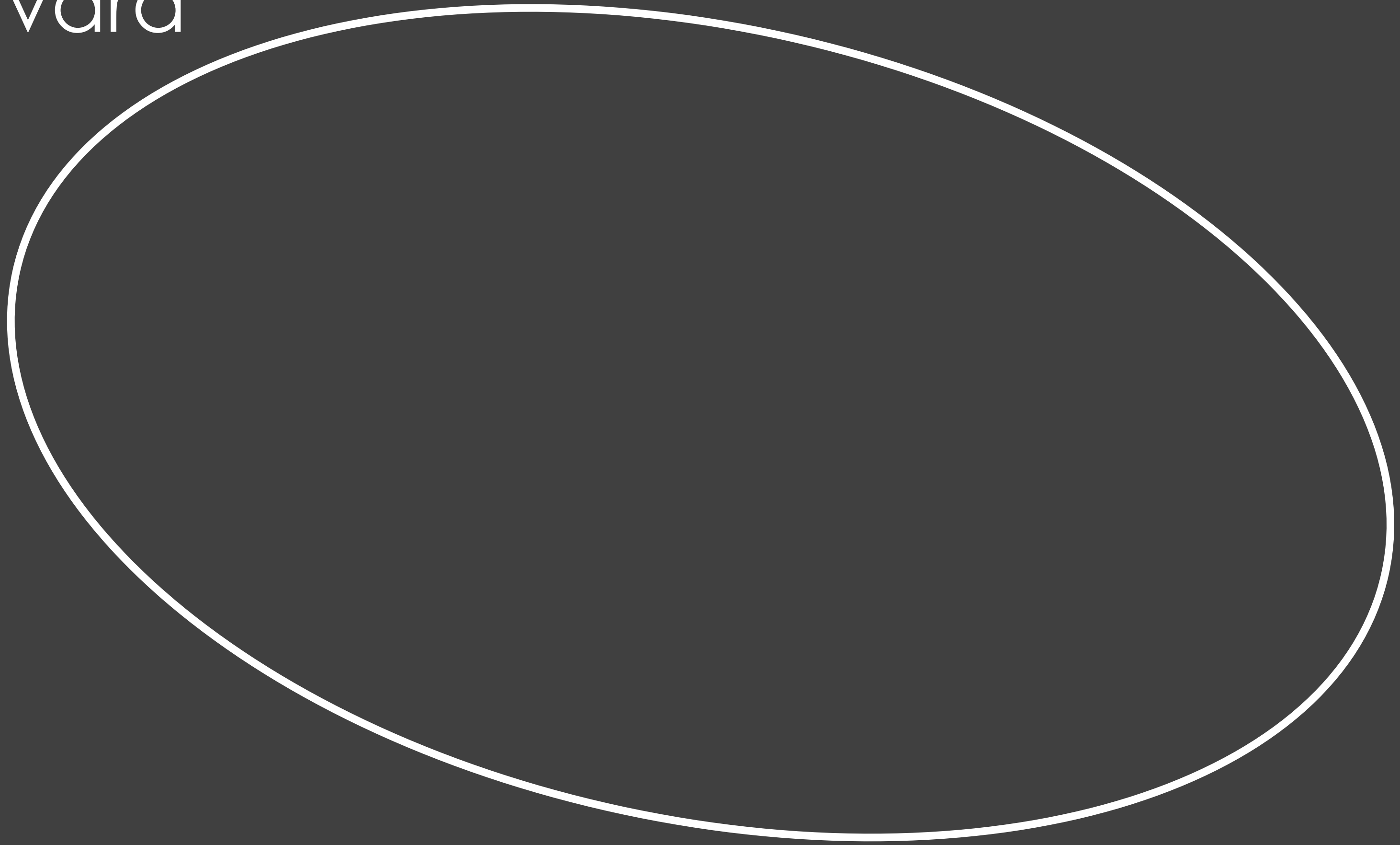
PEW RESEARCH CENTER

How might this impact the proportion of first born at Harvard?

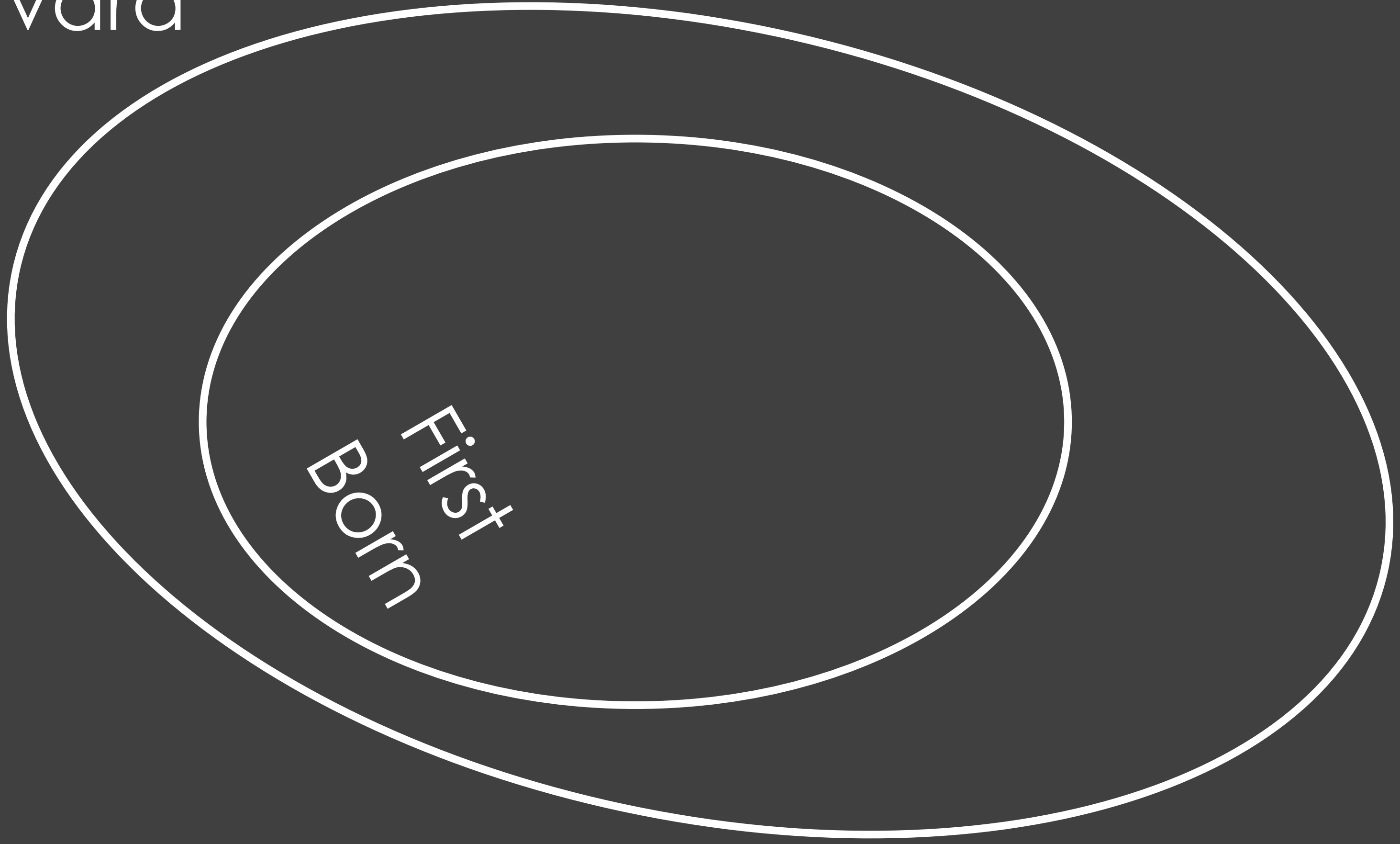
http://www.pewsocialtrends.org/2015/05/07/family-size-among-mothers/

Are first-borns more likely to attend Harvard?

Children of mothers
with at least one child
at Harvard

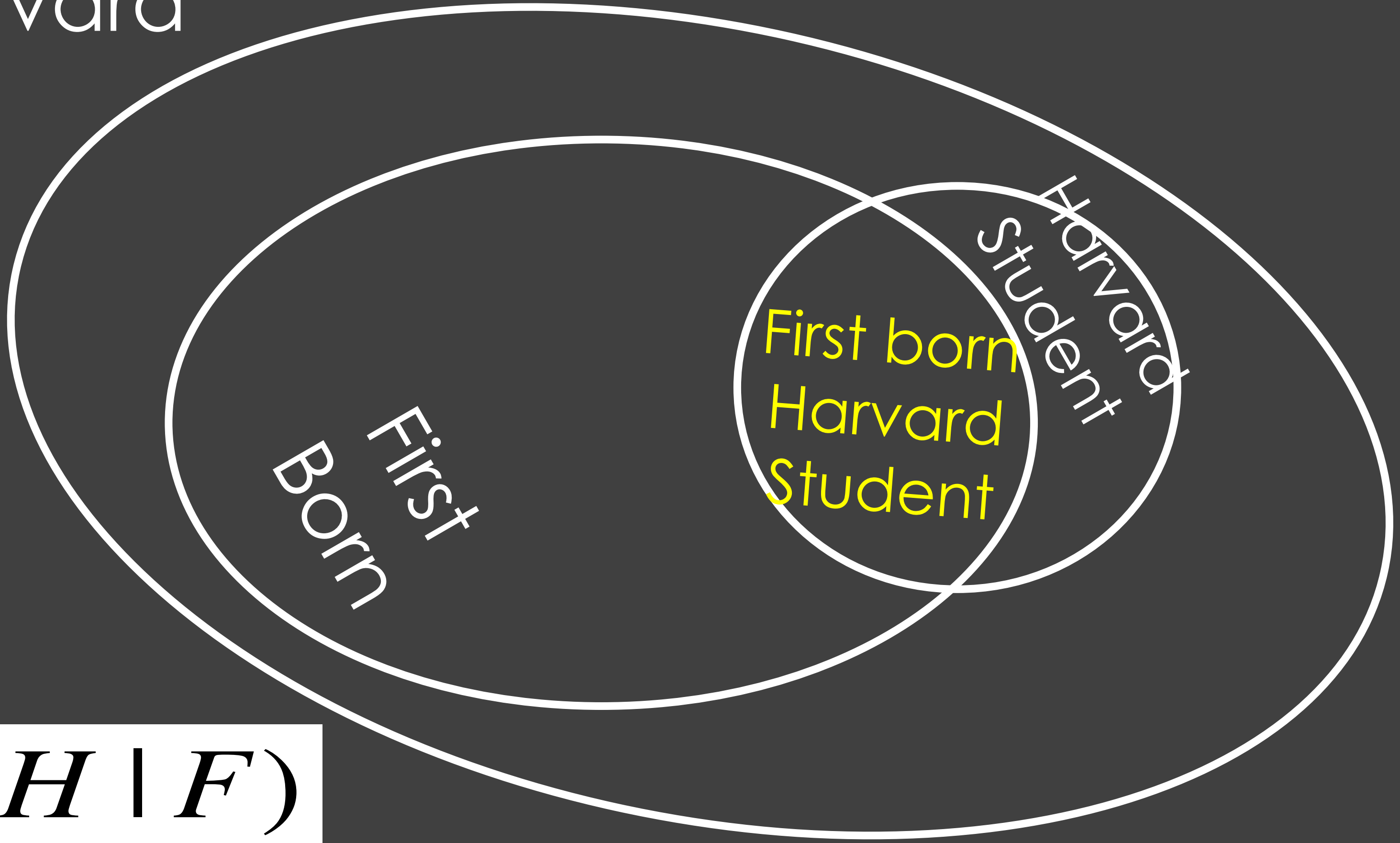


Children of mothers with at least one child at Harvard



First Born

Children of mothers with at least one child at Harvard



Want $P(H \mid F)$

Find the chance first born is at Harvard

We want to find the chance you are a Harvard student (H) given you are first born (F) and Compare it to the chance you are a Harvard student given you are not first born

$$P(H \mid F) = \frac{P(H \cap F)}{P(F)}$$

$$P(H \mid F^c) = \frac{P(H \cap F^c)}{P(F^c)}$$

Find the chance first born is at Harvard

Recall

$$P(H \cap F) = P(H)P(F \mid H)$$

Which implies Bayes Rule

$$P(H \mid F) = \frac{P(H)P(F \mid H)}{P(F)}$$

So the ratio becomes:

$$\frac{P(H \mid F)}{P(H \mid F^c)} = \frac{P(F \mid H)}{P(F^{c|} \mid H)} \times \frac{1 - P(F)}{P(F)}$$

Professor observed ¾ for P(F|H). But we need the fertility rate for Harvard moms

Find the chance first born is at Harvard

So the ratio becomes:

$$P(F) = \frac{1}{fertility}$$

$$r = \frac{P(H \mid F)}{P(H \mid F^c)} = \frac{3/4}{1/4} \times \frac{1 - P(F)}{P(F)} = 3(fertility - 1)$$

We need the fertility rate for Harvard moms…

If we plug in the US value for 1994 we get:
r = 3.9

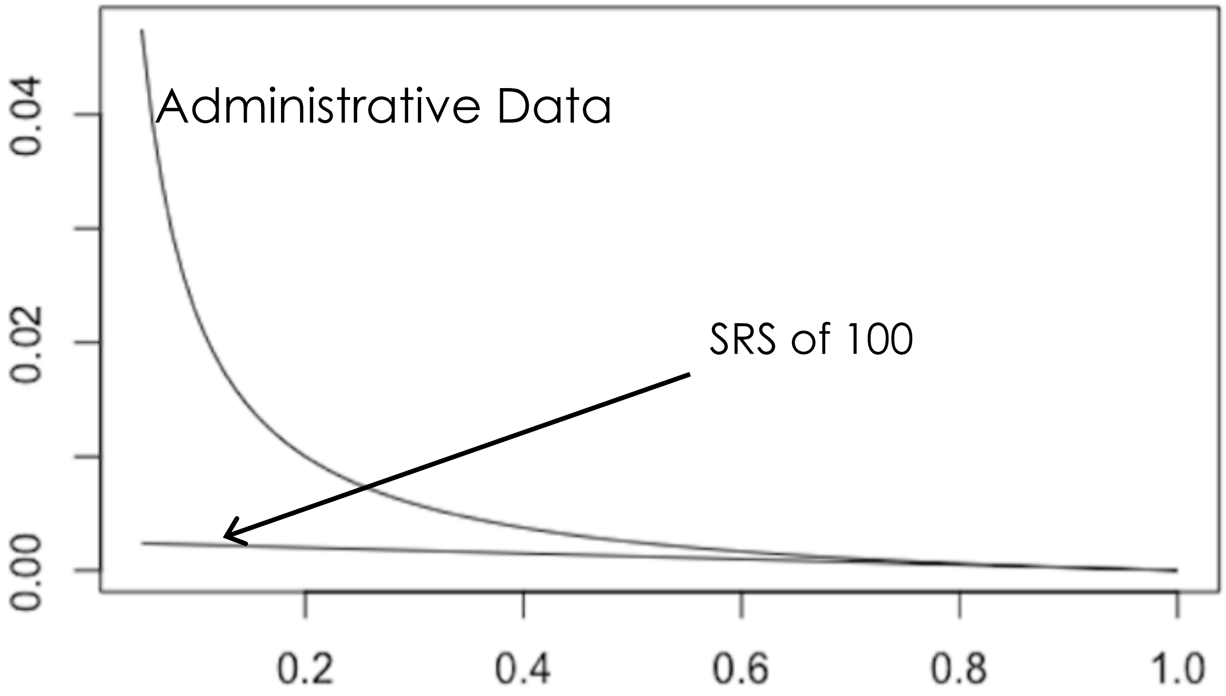
If we plug in the US highest educated value from Pew we get: r = 3.2

Can Big Data compensate for no SRS?

Big Data is **not** enough

With a slight correlation between the presence of the characteristic and the appearance in the dataset, the accuracy of a SRS of 100 outperforms a AD of 90%

I Got More Data, My Model Is More Refined…, Meng & Xie (2014) Econometrics Review


Administrative Data
SRS of 100