From COVID-19 Testing (and Vaccination) to Election Prediction: *How Small Are Our Big Data?*

Xiao-Li Meng

Harvard University

Suppose in Switzerland we randomly test 1000 people for COVID-19. How many people do we need to test (randomly) in US, which has about 40 times more people, to achieve the same statistical accuracy for estimating the positive rate in the US population?

- 1000
- 4000
- 10000
- 40000
- It depends

Why and when can we ignore the population size N?

- Think about tasting soup…
- Stir it well, then a few bits are sufficient

  regardless of the **size of the container**!

=

A Fundamental Identity for Statistical Estimation (Meng, 2018)

- X: Variable of interest (e.g., X=1 test positive; X=0 otherwise)
- R: Recording indicator (R=1, if a person is tested, R=0 otherwise)

$$\bar{x}_n - \bar{x}_N = \text{Corr}(X, R) \sqrt{\frac{N-n}{n}} \; \text{SD}(X)$$

Actual error = quality index × quantity index ×difficulty index

Definition: $\rho = \text{Corr}(X, R)$: **Data Defect Correlation**

How effective is selective testing for estimating infection rate for COVID-19?

- **Effective Sample Size** $\cong \frac{f}{1-f} \times \frac{1}{\rho^2}$, **where** f =n/N

- NY State: N=19.4 M, suppose we conduct ***n*=10,000 tests** (f =1/2000) and the selection effect is a **½ percent correlation** (ρ=0.005):

Same as conducting $\frac{0.0005}{0.9995} \times \frac{1}{0.005^2} \approx$ **20 random tests!**

(Walter Dempsey, Tweets, 4/5/2020)

- **A 99.80% loss of sample size due to the *Law of Large Population***

$$\bar{x}_n - \bar{x}_N = \frac{Ave(R_i X_i)}{Ave(R_i)} - Ave(X_i)$$

$$= \frac{Ave(R_i X_i) - Ave(R_i)Ave(X_i)}{Ave(R_i)} = \frac{Cov(R_i, X_i)}{Ave(R_i)}$$

$$= \frac{Cov(R_i, X_i)}{SD(R_i)SD(X_i)} \frac{SD(R_i)}{Ave(R_i)} SD(X_i)$$

$$= \text{Corr(X, R)} \sqrt{\frac{N-n}{n}} \text{ SD(X)}$$

Election Predictions

- Meng (2018) Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election. *Annals of Applied Statistics Vol 2: 685-726*

- Isakov and Kuriwaki (2020) Towards Principled Unskewing: Viewing 2020 Election Polls Through a Corrective Lens from 2016. *Harvard Data Science Review. https://doi.org/10.1162/99608f92.86a46f38*

Let's apply the formula to 2016 election

- **Using state-level survey data from Cooperative Congressional Election Study (CCES), conducted by**

  Stephen Ansolabehere, Brian Schaffner, Sam Luks,

  Douglas Rivers on Oct 4 - Nov 6, 2016 (YouGov);

- **Analysis assisted by Shiro Kuriwaki**

- Meng (2018) Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election. *Annals of Applied Statistics* Vol 2: 685-726

The Same Identity …

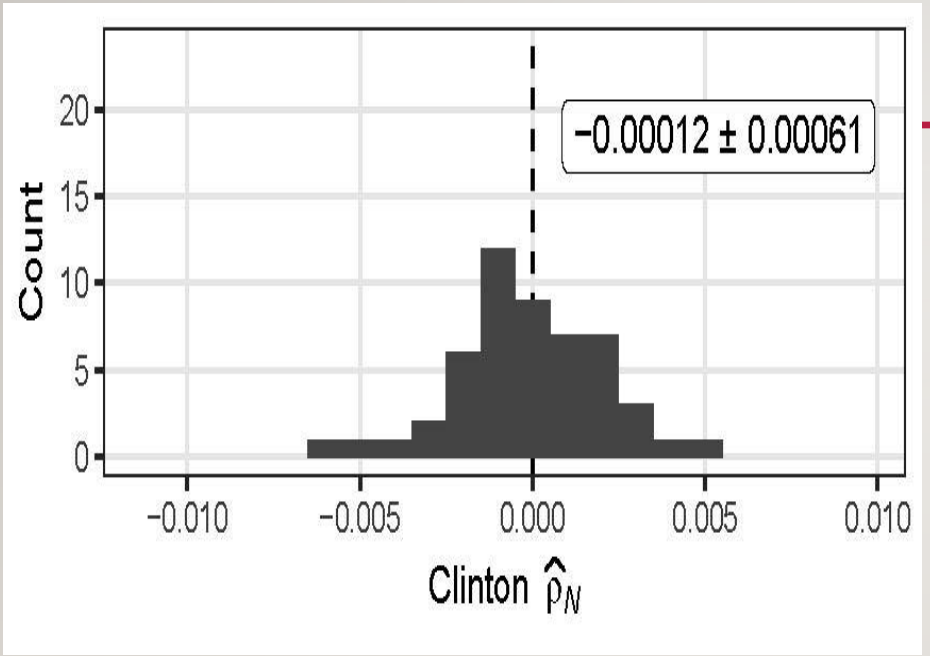
- X: variable of interest (e.g., X=1 vote for **A**; X=0 otherwise)
- R: Response indicator (R=1, if a person responds, R=0 otherwise)

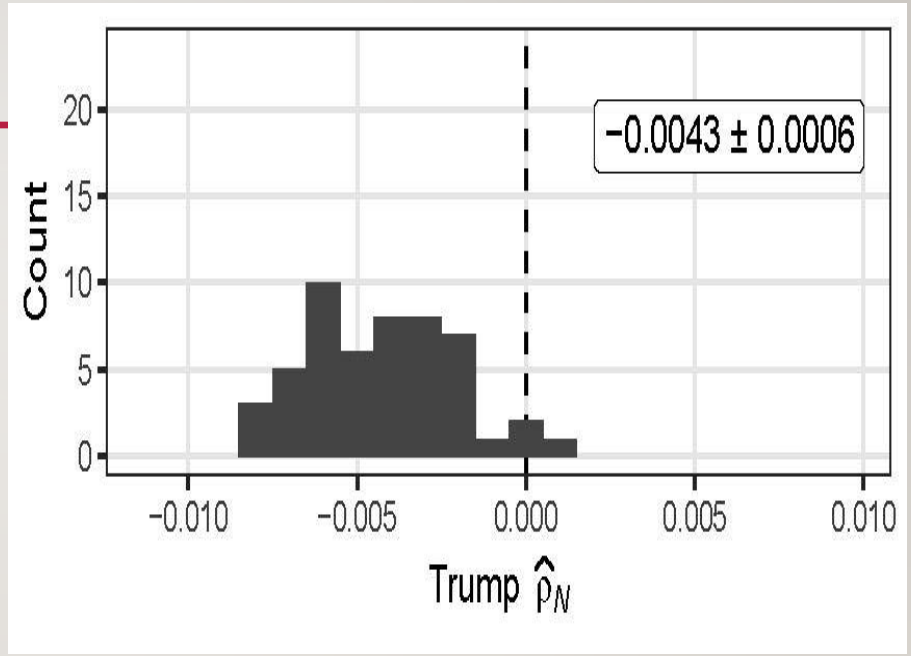
$$\bar{x}_n - \bar{x}_N = \text{Corr}(X, R) \sqrt{\frac{N-n}{n}} \; \text{SD}(X)$$

Actual error = quality index × quantity index ×difficulty index

Definition: $\rho = \text{Corr}(X, R)$: Data Defect Correlation

Converting the actual error to $\rho = \frac{\bar{x}_n - \bar{x}_N}{\sqrt{1-f}\,SD(X)} \frac{1}{\sqrt{N-1}} = \frac{Z}{\sqrt{N-1}}$





Clinton's ρ centered at 0

Trump's ρ centered at -0.005

What's the consequence of $\rho \simeq -0.005$?

- Given $\boldsymbol{\rho}$, the effective sample size of a "Big Data" with f=n/N, in terms of simple random sample size is (by matching the two MSEs)

$$\textbf{Effective Sample Size} \cong \frac{f}{1-f} \times \frac{1}{\rho^2}$$

- Take f=1%, n=2,300,000 for US voting population

(= 2,300 polls, each with 1000 people)

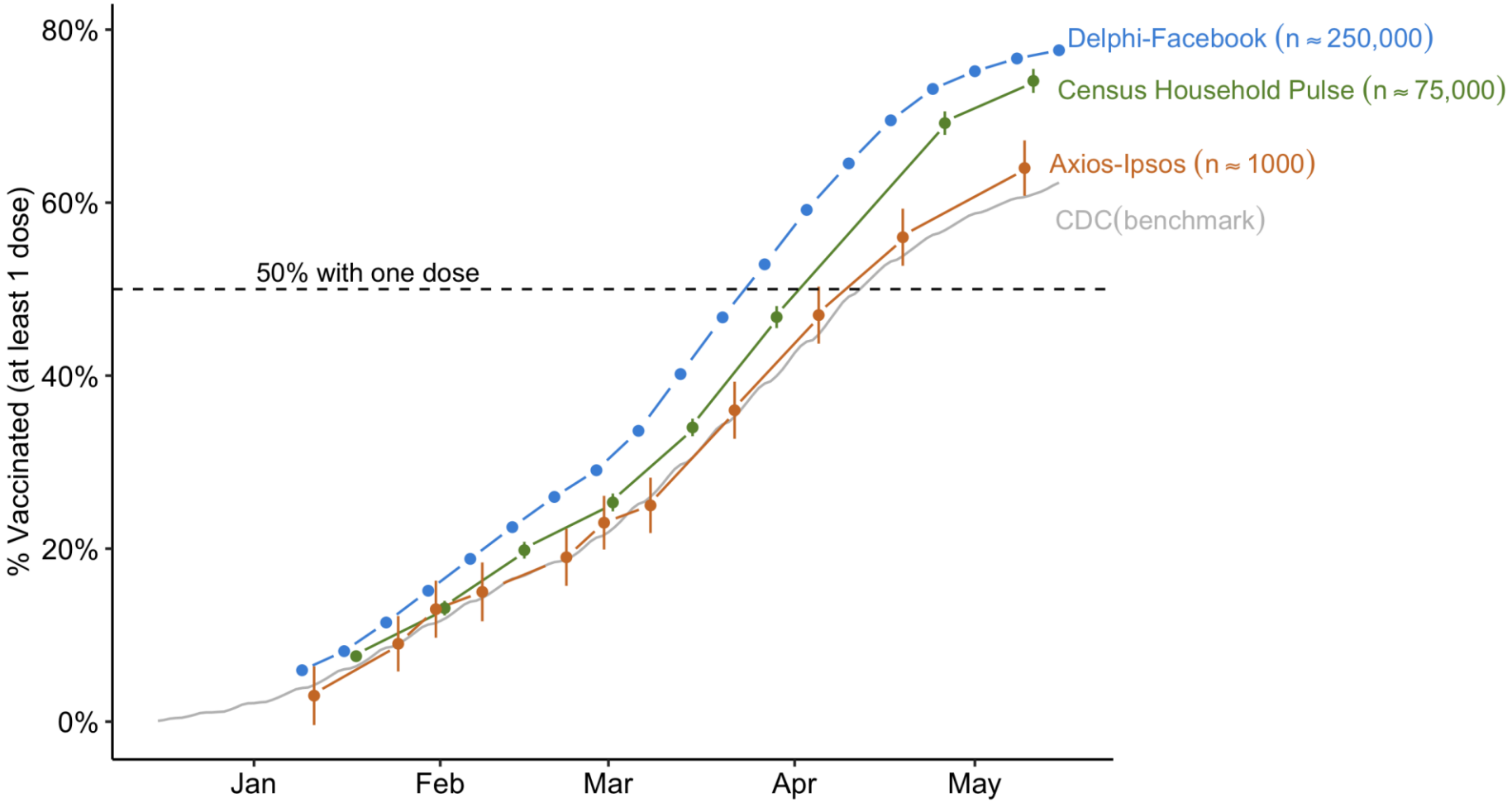
$$\textbf{Effective Sample Size} \cong \frac{0.01}{0.99} \times \frac{1}{0.005^2} \approx \mathbf{404}!$$

- ***A 99.98% lose of sample size -- How can we detect such problems in general, and before they happen?***

Covid-19 Vaccination

- Bradley, Kuriwaki, Isakov, Sejdinovic, Meng, Flaxman (2021) Are We There Yet? Big Data Significantly Overestimates COVID-19 Vaccination in the US. *arXiv: 2106.05818*

Estimating COVID-19 vaccine uptake



If we don't take data quality into account …

- **The Big Data Paradox:**

The Bigger the Data, the Surer We Fool Ourselves

- **The Law of Large Population**

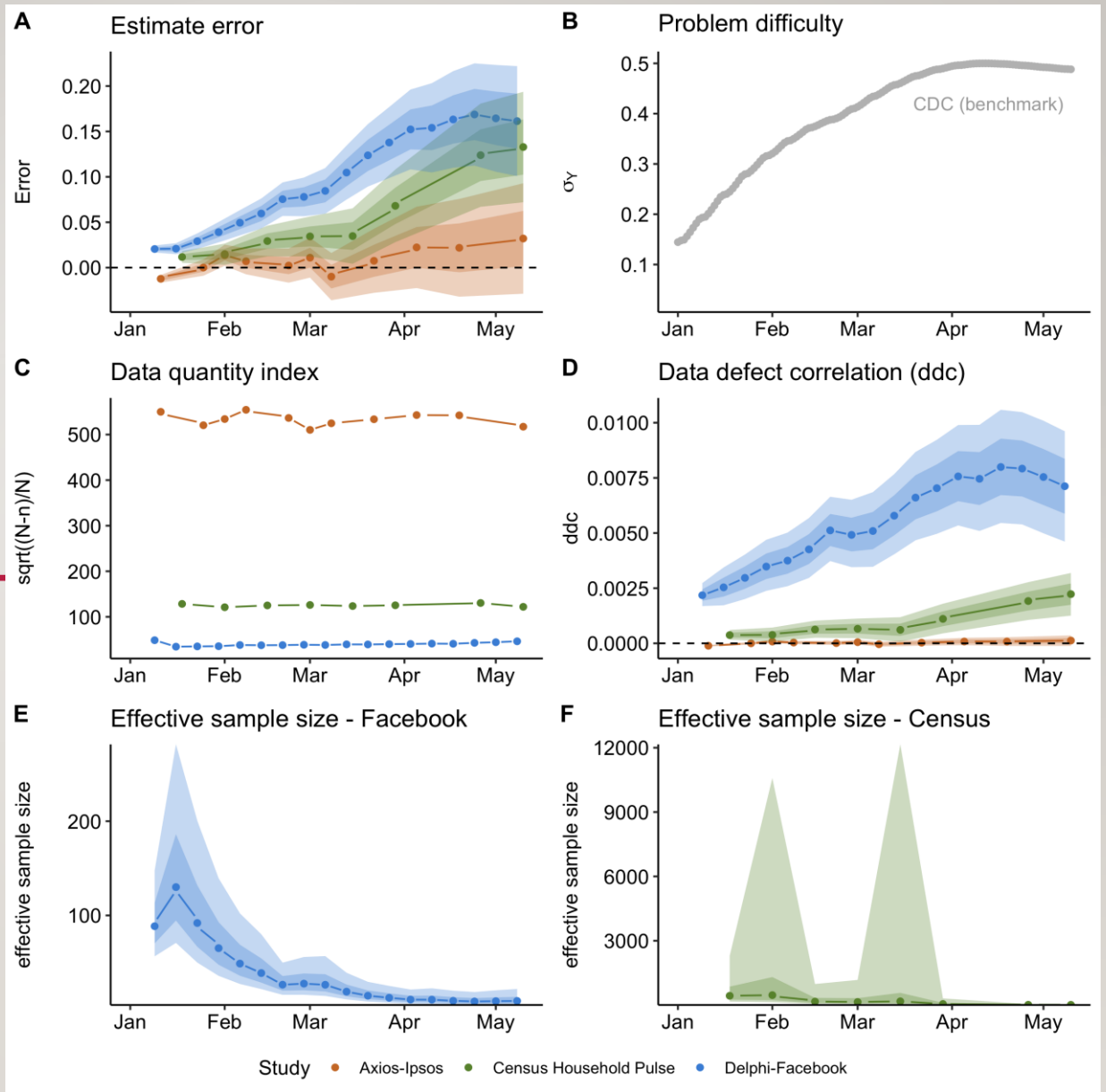
$$\text{Z score} = \rho\sqrt{N-1}$$

ρ: **Data Defect Correlation**

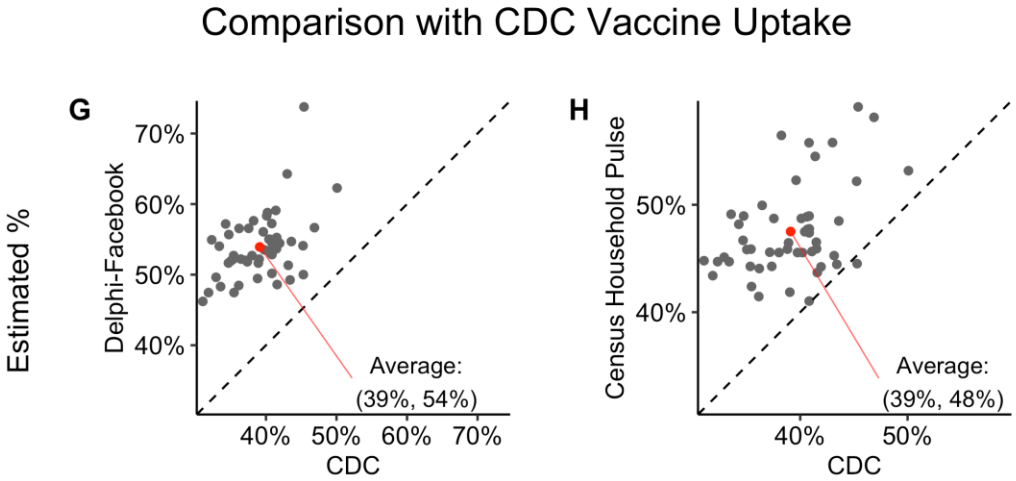
N: **Population size**

Decomposing the Estimation Error

- Quality index: ρ (d.d.c)
- Quantity index: $\sqrt{(N-n)/n}$
- Problem difficulty: SD(X)



Massachusetts is ranked 48th by both Face-book and Census surveys on vaccine uptake, but 7th by CDC – *Which one should the Mass governor trust (and act upon)?*



What lessons are learned from these two examples?

- In 2016 Election, 2.3 million underreported opinions is statistically equivalent to about 400 opinions without nonresponse, when the data defect correlation is -0.005
- For COVID-19, 250,000 biased sample form Facebook is statistical equivalent to no more than 250 random tests for estimating the vaccinee uptake in the US population, when ddc is in the range of 0.0025-0.0075.

Lessons learned

- Data quality is far more important than data quantity.
- Compensating for quality with quantity is a doomed game.
- It is far more important to reduce sampling and non-response biases than non-response rates.
- Invest in small but very high-quality surveys than large surveys with uncontrolled/unknown quality.
- When combining datasets, relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.
- For population inferences, "bigness" of "Big Data" should be measured by their relative size, not absolute size.

If we don't take data quality into account …

- **The Big Data Paradox:**

The Bigger the Data, the Surer We Fool Ourselves