Exploring the big data paradox for various estimands using vaccination data from the global COVID-19 Trends and Impact Survey (CTIS)

Youqi Yang
Co-Authors: Walter Dempsey, Peisong Han, Yashwant Deshmukh, Sylvia Richardson, Brian Tom, Bhramar Mukherjee

Department of Biostatistics, University of Michigan

August 6, 2024

Outline

1. Introduction
   - Types of sampling
   - Big data paradox
   - Study aims
2. Materials and methods
   - Data source
   - Method
3. Results
   - Big data paradox in India
   - Big data paradox beyond the mean
4. Discussions

Types of sampling

There are two primary types of sampling methods (Little, 2014).

Probability sampling

Every individual in the population has a **positive** probability of selection and every sample of a given size has a **known** probability of selection. It is required for design-based inference. It is subject to non-response bias.

Non-probability sampling

The probability of a given sample to be selected from the target population is **unknown**. It is more convenient and less costly. It is subject to both selection bias and non-response bias.

Big data paradox

Meng (2018) provided the following decomposition formula. Let n denote the sample size coming from a finite population of size N. Let $\overline{Y}_n$ be the sample average of the variable Y, and $\overline{Y}_N$ be the population average. Let the binary indicator R take the value 1 if we have recorded a value of Y in the sample and 0 otherwise.

$$\underbrace{\overline{Y}_n - \overline{Y}_N}_{\text{Estimation error}} = \underbrace{\rho_{Y,R}}_{\text{Data defect correlation}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data deficiency}} \times \underbrace{\sigma_Y}_{\text{Inherent problem difficulty}}.$$

Big data paradox

$$\underbrace{\overline{Y}_n - \overline{Y}_N}_{\text{Estimation error}} = \underbrace{\rho_{Y,R}}_{\text{Data defect correlation}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data deficiency}} \times \underbrace{\sigma_Y}_{\text{Inherent problem difficulty}}.$$

- Advantage: compare surveys with different recording mechanisms and sample sizes drawn from the same population for estimating the same target population quantity.
- We can also calculate the bias-adjusted effective sample size (n_{eff}) by,

$$n_{eff} = \frac{n}{N-n} \times \frac{1}{\hat{\rho}^2_{Y,R}}.$$

It is defined as the size of a simple random sample (SRS) drawn from the same population that will produce the same mean squared error (MSE) as observed in the survey of interest.

Big data paradox

$$\underbrace{\overline{Y}_n - \overline{Y}_N}_{\text{Estimation error}} = \underbrace{\rho_{Y,R}}_{\text{Data defect correlation}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data deficiency}} \times \underbrace{\sigma_Y}_{\text{Inherent problem difficulty}}.$$

- However, to estimate $\rho_{Y,R}$, we need to know $\overline{Y}_N$ and σ_Y.
- A rare and unique situation: Using reports from the US Centers for Disease Control and Prevention (CDC) as the benchmark, Bradley et al. (2021) found a big non-probability sample (average weekly sample size $\sim$ 250,000; the COVID-19 Trends and Impact Survey, CTIS) produced a more biased estimate of the first-dose vaccination rate among US adults when compared to a small probability survey (average weekly sample size $\sim$ 1,000; the Axios-Ipsos survey) but with narrower confidence intervals. This is known as the **big data paradox**.

Study aims

1. **Big data paradox in India:** To check whether the Big Data Paradox that has been explored in the US also holds for a populous country like India with a completely different fabric of society and healthcare system.
2. **Big data paradox beyond the mean:** To check whether a survey that is substantially biased in estimating the population average of the target variable could still be useful if the estimand is changed to (a) the successive difference or relative difference in the population average over time; (b) differences in population averages of two subgroups.

Data source

CTIS is conducted jointly by Meta/Facebook along with Carnegie Mellon University (the US) and the University of Maryland (the globe including India).

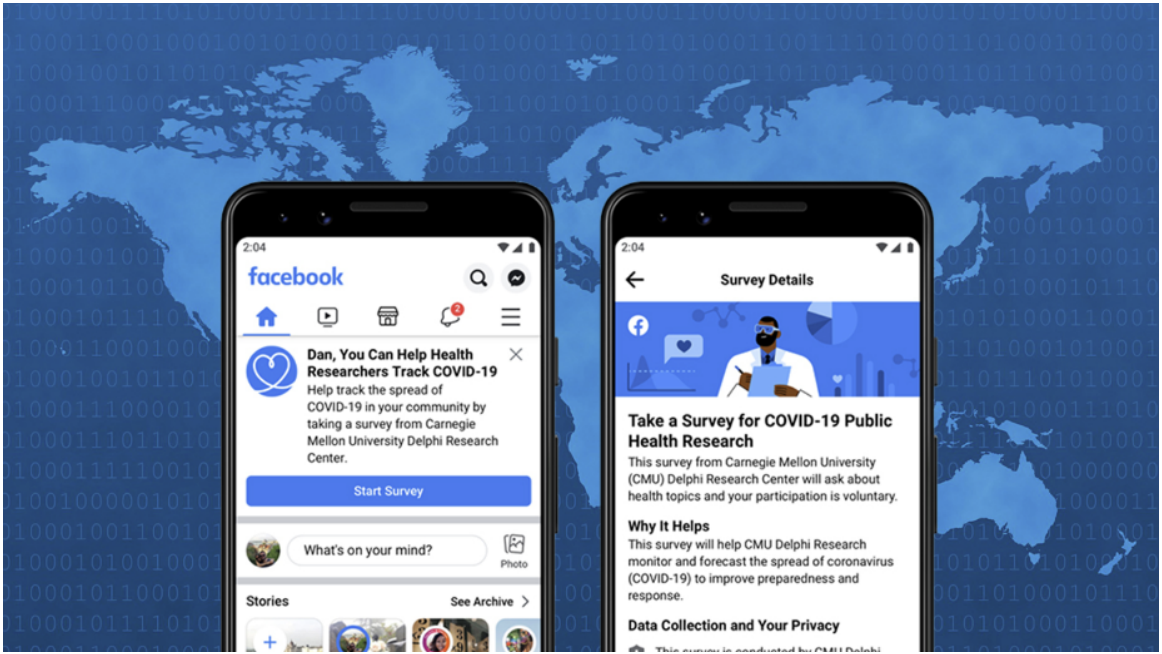

Figure 1: An example of the CTIS survey.

Data source

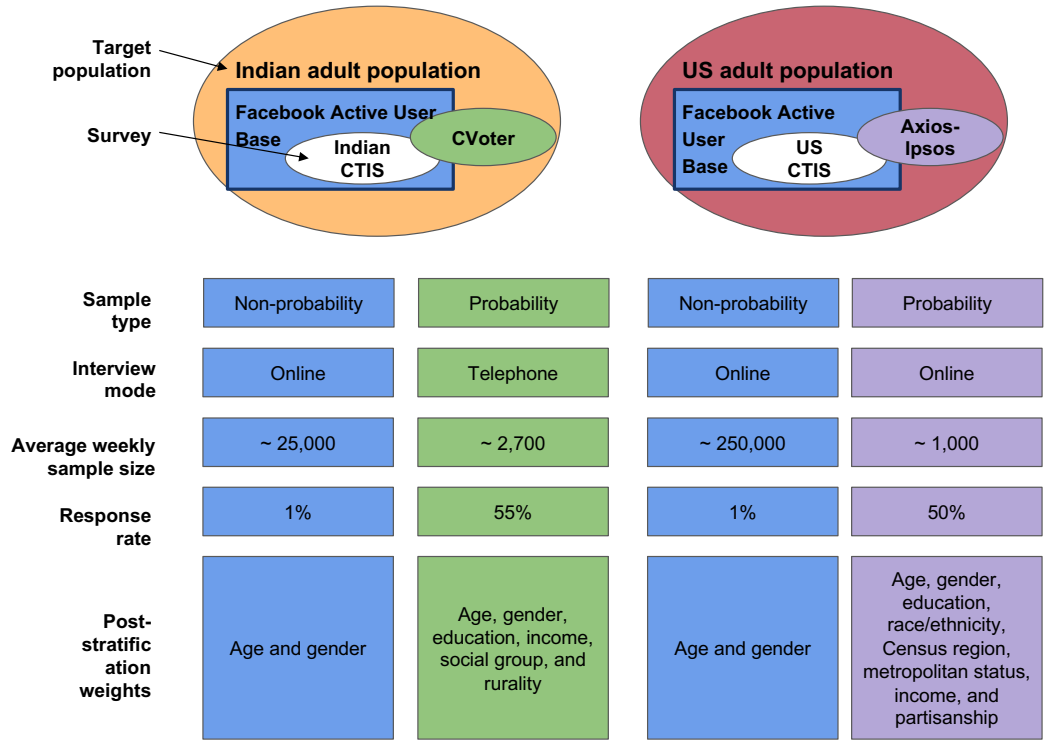

Figure 2: Data sources.

Successive (relative) successive difference in means

Let $\overline{Y}_{n,t}$ be the survey average of the variable Y at time t and $\overline{Y}_{N,t}$ be the population average at the same time t. Let $\sigma^2_{Y_t}$ be the population variance of Y at time t. Define the effective sample size n_{eff} as the size of two SRS samples that can generate the observed MSE for the successive difference and relative successive difference.

$$\text{Difference}(n_{eff}) = \frac{\sigma^2_{Y_{t-1}} + \sigma^2_{Y_t}}{[(\overline{Y}_{n,t} - \overline{Y}_{n,t-1}) - (\overline{Y}_{N,t} - \overline{Y}_{N,t-1})]^2}.$$

$$\text{Relative Difference}(n_{eff}) = \frac{\left(\frac{\overline{Y}_{N,t}}{\overline{Y}_{N,t-1}}\right)^2 \left[\frac{\sigma^2_{Y_{t-1}}}{(\overline{Y}_{N,t-1})^2} + \frac{\sigma^2_{Y_t}}{(\overline{Y}_{N,t})^2}\right]}{\left[\frac{\overline{Y}_{n,t} - \overline{Y}_{n,t-1}}{\overline{Y}_{n,t-1}} - \frac{\overline{Y}_{N,t} - \overline{Y}_{N,t-1}}{\overline{Y}_{N,t-1}}\right]^2}.$$

Subgroup difference in means

Let $\overline{Y}_{n_g}$ be the survey average of the variable Y in the subgroup g, and let $\overline{Y}_{N_g}$ be the population average in the same subgroup. Let n_g be the sample size of the survey conducted in the subgroup g and N_g be the population size of the subgroup g. For simplicity, we assume the general population can be explicitly divided into two partitions: group I and group II, $g = \mathrm{I}, \mathrm{II}$. Denote a binary indicator G taking 1 if a response belongs to group II and 0 otherwise. Consider a new variable $Y^* = Y \times G - Y \times (1 - G)$ and denote its binary indicator as R^*.

$$(\overline{Y}_{n_{\mathrm{II}}} - \overline{Y}_{n_{\mathrm{I}}}) - (\overline{Y}_{N_{\mathrm{II}}} - \overline{Y}_{N_{\mathrm{I}}}) = \rho_{Y^*,R^*} \times \sqrt{\frac{(N_{\mathrm{I}} + N_{\mathrm{II}}) - (n_{\mathrm{I}} + n_{\mathrm{II}})}{n_{\mathrm{I}} + n_{\mathrm{II}}}} \times \sigma_{Y^*}.$$

$$n_{eff} = \frac{n_{\mathrm{I}} + n_{\mathrm{II}}}{(N_{\mathrm{I}} + N_{\mathrm{II}}) - (n_{\mathrm{I}} + n_{\mathrm{II}})} \times \frac{1}{(\hat{\rho}_{Y^*,R^*})^2}.$$

Results: Big data paradox in India

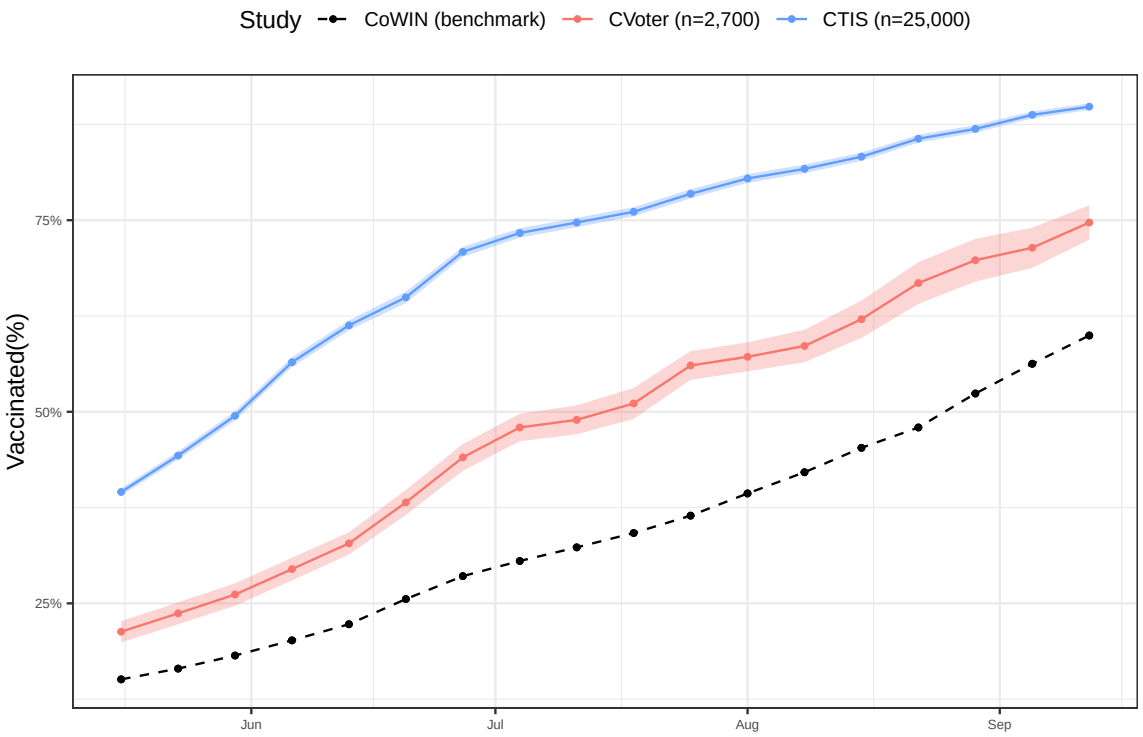

Figure 3: Vaccine uptake in India.

Results: Big data paradox in India

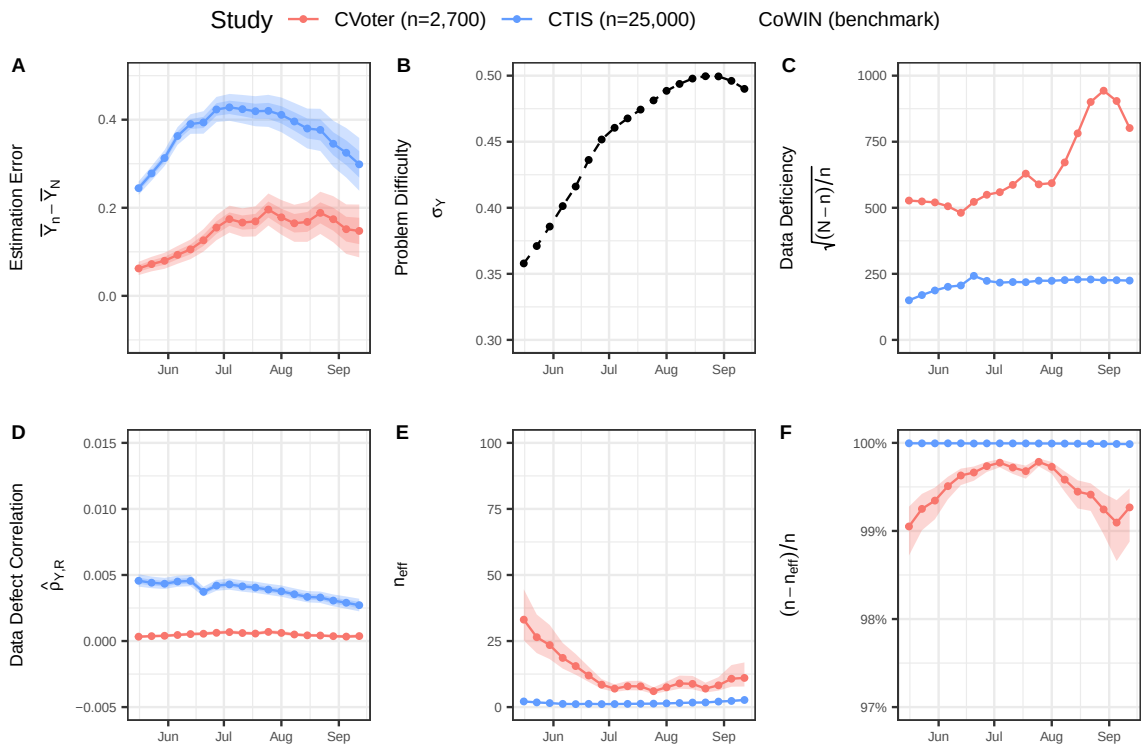

Figure 4: Decomposition of the estimation error and the effective sample size of vaccine uptake in India.

Results: Big data paradox in successive difference

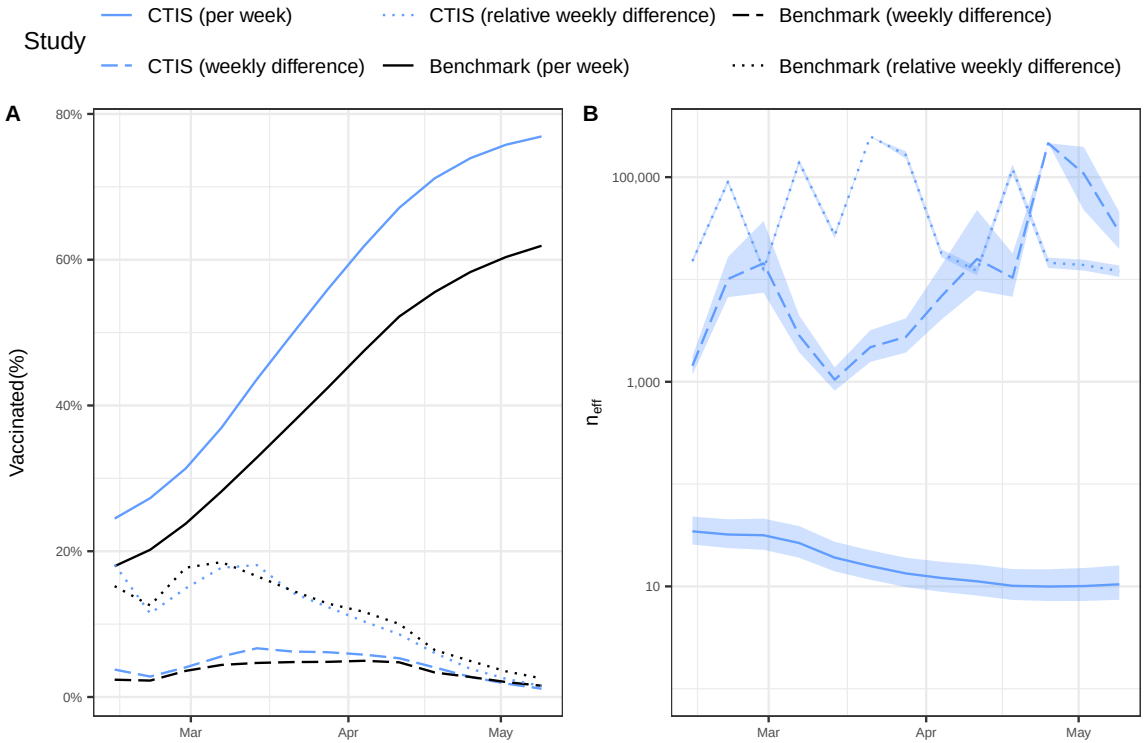

Figure 5: The successive difference and relative successive difference of vaccine uptake in the US.

Results: Big data paradox in subgroup difference

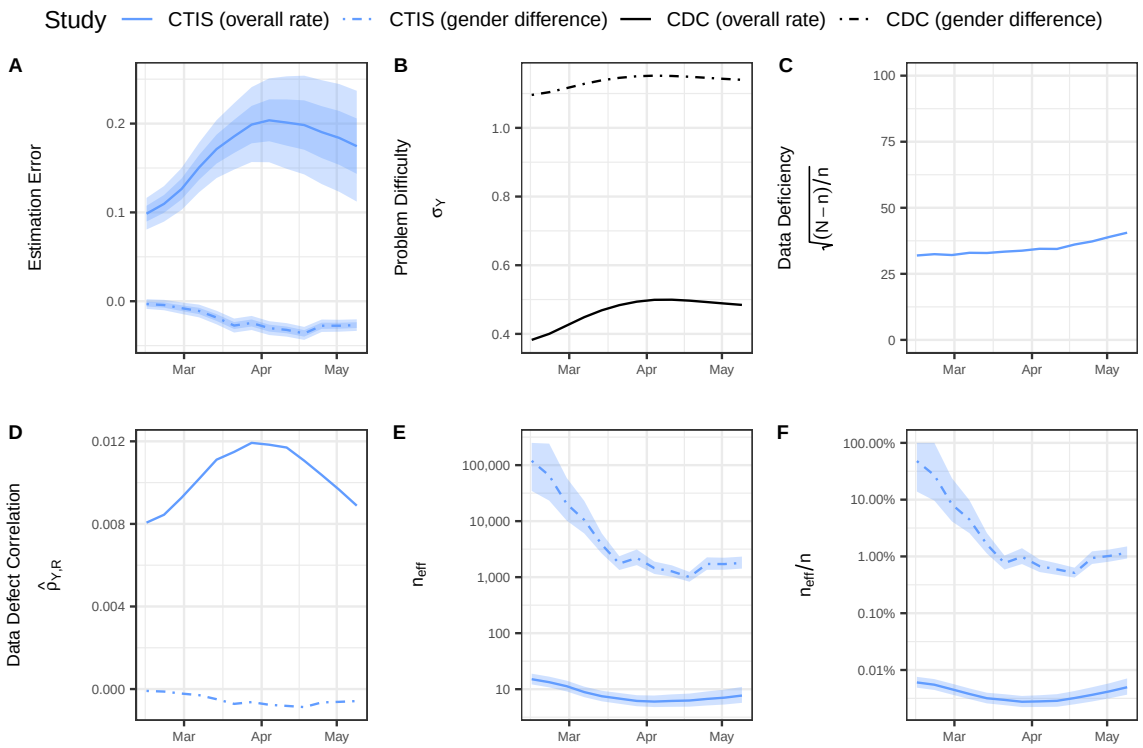

Figure 6: Decomposition of the estimation error and effective sample size of gender difference in vaccine uptake among US adults.

Discussions

- A large non-probability sample (namely CTIS) produces more biased estimates of vaccination uptake among Indian adults in comparison to a small probability survey (namely CVoter). This emphasizes that the **big data paradox** holds data from countries beyond the US.
- The US CTIS can provide more accurate estimates of (a) the successive difference and relative successive difference in vaccination rates, and (b) gender differences in vaccination rates. These offer a more optimistic assessment of the CTIS data and similar large non-probability samples, indicating that the **big data paradox** is not inevitable for every estimand one may be interested in.

Discussions

- However, the small probability survey (namely CVoter) did not estimate the overall vaccination rates that closely matched the benchmark.

$$I(\text{Response is recorded for a participant}) =$$

$$I(\text{Participant included in the study})$$

$$\times I(\text{Response is recorded}|\text{Participant is included}).$$

In a probability survey, although we have knowledge of the first component, the non-response mechanism remains a major factor driving selection bias.

Acknowledgments

We would like to express our sincere gratitude to:
Social Data Science Center at the University of Maryland
Delphi Group at Carnegie Mellon University
Professor F. Kreuter
Professor X.-L. Meng

Our paper can be accessed through: Yang, Y., Dempsey, W., Han, P., Deshmukh, Y., Richardson, S., Tom, B., and Mukherjee, B. (2024). Exploring the big data paradox for various estimands using vaccination data from the global COVID-19 Trends and Impact Survey (CTIS). *Science Advances, 10*, eadj0266.
https://doi.org/10.1126/sciadv.adj0266

An R shiny app for the big data paradox in 85 additional countries using CTIS:
https://3ogdqc-youqi-yang.shinyapps.io/LLPinVaccine/.

References

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600(7890):695–700.

Little, R. J. (2014). Survey sampling: Past controversies, current orthodoxy, and future paradigms. *Past, present, and future of statistical science*, pages 413–428.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2).

Thank you! Questions?