HDSR

HARVARD DATA SCIENCE REVIEW

A Telescopic, Microscopic, and Kaleidoscopic View of Data Science

2021 Best New Journal in Science, Technology, & Medicine

As an open access platform of the Harvard Data Science Initiative, the *Harvard Data Science Review* features **foundational thinking, research milestones, educational innovations, and major applications,** with a primary emphasis on reproducibility, replicability, and readability. It aims to publish contents that help to define and shape data science as a scientifically rigorous and globally impactful multidisciplinary field based on the principled and purposed production, processing, parsing and analysis of data. By uniting the strengths of a premier research journal, a cutting-edge educational publication, and a popular magazine, HDSR provides a crossroads at which fundamental data science research and education intersect directly with societally-important applications from industry, governments, NGOs, and others. By disseminating inspiring, informative, and intriguing articles and media materials, HDSR aspires to be a global forum on ***everything data science and data science for everyone.***

Issue 1.1, Summer 2019 ••• 2 more Published on Jul 01, 2019 DOI 10.1162/99608f92.ba20f892

Data Science: An Artificial Ecosystem

Issue 1.1 / Summer 2019

by Xiao-Li Meng

Published on Jul 01, 2019

last released
4 months ago

Type the word 'scientist' into your favorite browser and search for images. Most likely you will see photos of actual scientists from various fields. Now repeat the search, using 'data scientist.' You will see far fewer photos, but many animated figures standing by or pointing to various lists of must-have skills that read like tiger parents' assignments for their children.

As we advance deeper into the digital age, our societal demands for data scientists naturally rise in both quantity and quality. Most of us have some knowledge about other kinds of scientists (granted that such knowledge can be quite flawed), but we are much less clear about who data scientists are and what they do. Indeed, what exactly is *data science* (DS)? As you may have guessed, the answer depends on whom you ask. Some say DS is CS (computer science). Others think DS is simply S (statistics). You may even run into someone who declares DS is just hyped-up BS (and I don't mean "Bayesian statistics").

Generation–Collection–Processing–Storage–Management–Analysis–Visualization–Interpretation

Data Conceptualization

Data governance is key to interpretation: Reconceptualizing data in data science

Sabina Leonelli
Sociology, Philosophy and Anthropology,
University of Exeter

The Data Life Cycle

Jeannette M. Wing
Computer Science
Columbia University

Data Curation and Provenance

The Lives and After Lives of Data

Christine L. Borgman
Information Studies, UCLA

Data Science for Governing and Policy Making

An Interview With Murray Edelman on the History of the Exit Poll

by Murray Edelman, Liberty Vittert, and Xiao-Li Meng

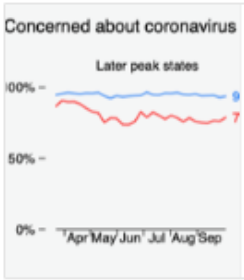
An interview with Murray Edelman by Liberty Vittert and Xiao-Li Meng

Understanding the Quality of the 2020 Census Is Essential

by John Thompson

Recombination: A Family of Markov Chains for Redistricting

by Daryl DeFord, Moon Duchin, and Justin Solomon



The Politics of COVID-19: Partisan Polarization About the Pandemic Has Increased, but Support for Health Care Reform Hasn't Moved at All

by John Sides, Chris Tausanovitch, and Lynn Vavreck

SCRAM: A Platform for Securely Measuring Cyber Risk

by Leo de Castro, Andrew W. Lo, Taylor Reynolds, Fransisca Susan, Vinod Vaikuntanathan, Daniel Weitzner, and Nicolas Zhang

Error, Uncertainty, and the Shifting Ground of Census Data

by Dan Bouk

Implementing Differential Privacy: Seven Lessons From the 2020 United States Census

by Michael B. Hawes

Measuring the Gross Domestic Product (GDP): The Ultimate Data Science Project

by Brian C. Moyer and Abe Dunn

2022/11/10

5

Differential Privacy for the 2020 U.S. Census: Can We Make Data Both Private and Useful?

Special Issue 2

FROM THE EDITORS

Harnessing the Known Unknowns: Differential Privacy and the 2020 Census

by Ruobin Gong, Erica L. Groshen, and Salil Vadhan

Published: Jun 24, 2022

Special Issue 2: Differential Privacy for the 2020 U.S. Census

CENSUS: IMPORTANCE, HISTORY, AND TECHNICAL CHANGES

Data Science for Governing and Policy Mak ••• 3 more Published on Jan 31, 2020 DOI 10.1162/99608f92.c871f9e0 SHOW DETAILS

Coming to Our Census: How Social Statistics Underpin Our Democracy (and Republic)

by Teresa A. Sullivan

Published on Jan 31, 2020

last released 7 months ago

ABSTRACT

The 2020 Census provides the opportunity to reflect on the key role statisticians, demographers, and other social scientists play in safeguarding American democracy. Democracy requires numbers for its proper functioning, and there is now a large statistical infrastructure of which the constitutionally mandated census is the keystone. Mistrust of the government is a major obstacle for the census, potentially affecting both accuracy and completeness. The mistrust is stimulated by fears of individual or household census data being willingly or inadvertently shared with other government agencies (data privacy issues) or even foreign actors (hacking). As two 2019 Supreme Court decisions in juxtaposition suggest, no checks or balances protect the integrity of the census. The professional integrity of statisticians is the best defense of the census.

Keywords: census, democracy, statistical infrastructure, data errors, data privacy

This article is accompanied by multiple invited discussion pieces and a rejoinder by the author.





How my own research has been affected/evolved …

- Data Quality (Multi-source Inference/Learning)
- Data Privacy (Multi-phase Inference/Learning)
- Data Granularity (Multi-resolution Inference/Learning)

EVERYTHING DATA SCIENCE & DATA SCIENCE FOR EVERYONE

INAUGURAL VOLUME | 2019
HDSR.MITPRESS.MIT.EDU

HARVARD DATA SCIENCE REVIEW

What does the educated citizen need to know about Data Science?

What is a data life cycle?

Are we in an AI revolution?

Is there a data creators' advantage?

Who wrote "In My Life"?

What is the Evidence Act?

A Conversation With Steve Ballmer

Building Representative Miniatures out of Non-representative Big Data: An Interplay of Data Quantity, Quality, and Privacy

Xiao-Li Meng, Harvard University

Building Representative Miniatures out of Non-representative Big Data: An Interplay of Data Quantity, Quality, and Privacy

Xiao-Li Meng, Harvard University

- Meng (2022). Miniaturizing Data Defect Correlation: A Versatile Strategy for Handling Non-Probability Samples. *Survey Methodology*.
- Meng (2018) **Statistical Paradises and Paradoxes in Big Data (I): From The Law of Large Populations to The Big Data Paradox.** *Annals of Applied Statistics*, 12, 685–726.
- Bradley, Sejdinovic, Meng, Kuriwaki, Isakov, Flaxman (2021) **Unrepresentative Big Surveys Significantly Overestimated COVID-19 Vaccination in the US.** *Nature* Dec; 600(7890):695-700.

It has been a decade ...

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g., bias, predictive power).

It has been a decade ...

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g., bias, predictive power).
- **But is an 80% non-random sample "better" than a 5% random sample in measurable terms? 90%? 95%? 99%?** (Jeremy Wu of US Census Bureau, 2012, Seminar at Harvard Statistics)

It has been a decade ...

- We know that a 5% random sample is better than a 5% non-random sample in measurable ways (e.g., bias, predictive power).
- **But is an 80% non-random sample "better" than a 5% random sample in measurable terms? 90%? 95%? 99%?** (Jeremy Wu of US Census Bureau, 2012, Seminar at Harvard Statistics)
- **"Which one should we trust more: a 1% survey with 60% response rate or a non-probabilistic dataset covering 80% of the population?"** (Keiding and Louis, 2015, Joint Statistical Meetings; and *JRSSB*, 2016)

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

- Landmark paper: Jerzy Neyman (1934)

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

- Landmark paper: Jerzy Neyman (1934)
- The "revolution" lasted about 50 years (Jelke Bethlehem, 2009)

It has been a few centuries ..

- **Law of Large Numbers**:
  Jakob Bernoulli (1713)
- **Central Limit Theorem**:
  Abraham de Moivre (1733):
  $\text{error} \propto \frac{1}{\sqrt{n}}$: $n - \text{sample size}$
- **Survey Sampling**:
  - Graunt (1662); Laplace (1882)
  - The "**intellectually violent revolution**" in 1895 by Anders Kiær, Statistics Norway

- Landmark paper: Jerzy Neyman (1934)
- The "revolution" lasted about 50 years (Jelke Bethlehem, 2009)
- First implementation in US Census: 1940 led by Morris Hansen

Triumphs and Troubles of Probability Sampling

Triumphs: Ensuring learning is feasible and cost effective

Triumphs and Troubles of Probability Sampling

Triumphs: Ensuring learning is feasible and cost effective

- Census is not feasible in most studies
- Sampling makes a study feasible and even cost effective
- Sampling also reduces privacy cost since it is a **random suppression**

Triumphs and Troubles of Probability Sampling

Triumphs: Ensuring learning is feasible and cost effective

- Census is not feasible in most studies
- Sampling makes a study feasible and even cost effective
- Sampling also reduces privacy cost since it is a **random suppression**

Troubles: Ensuring "repetitiveness" is increasingly challenging

Triumphs and Troubles of Probability Sampling

Triumphs: Ensuring learning is feasible and cost effective

- Census is not feasible in most studies
- Sampling makes a study feasible and even cost effective
- Sampling also reduces privacy cost since it is a **random suppression**

Troubles: Ensuring "repetitiveness" is increasingly challenging

- Non-response/coverage bias (not merely the rate) is on the rise
- Social media data are non-representative by design
- Administrative data are not probability samples

Thoughts: Handling data defects and privacy concomitantly

Triumphs and Troubles of Probability Sampling

Triumphs: Ensuring learning is feasible and cost effective

- Census is not feasible in most studies
- Sampling makes a study feasible and even cost effective
- Sampling also reduces privacy cost since it is a **random suppression**

Troubles: Ensuring "repetitiveness" is increasingly challenging

- Non-response/coverage bias (not merely the rate) is on the rise
- Social media data are non-representative by design
- Administrative data are not probability samples

Thoughts: Handling data defects and privacy concomitantly

- Can we utilize the inherent data defects as data privacy protections?
- Can we engineer "representative miniatures" that trade data utility and privacy more sensibly than adding noises to defective data?

Statistical Thinking with R ...

Statistical Thinking with R ...

Recording (Reporting/Responding) Indicator R for a finite population $\mathbb{X} = \{X_1, \ldots, X_N\}$:

$R_i = 1$, if X_i is included in a sample of size n, and $R_i = 0$ otherwise

Statistical Thinking with R ...

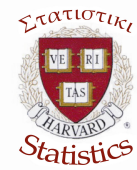
Recording (Reporting/Responding) Indicator R for a **finite population** $\mathbb{X} = \{X_1, \ldots, X_N\}$:

$R_i = 1$, if X_i is included in a sample of size n, and $R_i = 0$ otherwise

Design Prob: Prob controlled and implemented by human

Ex: Prob sampling, $\Pr(R_i = 1|\mathbb{X}) = n/N$

More: Bootstraps, clinical trials, differential privacy, permutation test, MC.



Statistical Thinking with R ...

Recording (Reporting/Responding) Indicator R for a finite population $\mathbb{X} = \{X_1, \ldots, X_N\}$:

$R_i = 1$, if X_i is included in a sample of size n, and $R_i = 0$ otherwise

Design Prob: Prob controlled and implemented by human

Ex: Prob sampling, $\Pr(R_i = 1|\mathbb{X}) = n/N$

More: Bootstraps, clinical trials, differential privacy, permutation test, MC.

Divine Prob: Existential Prob by God/Nature or by faith.

Ex: Real-World Evidence, but we assume R is random, and $\Pr(R_i = 1|\mathbb{X}) = \pi(X_i)$. [Needed for define "missing at Random"]

Statistical Thinking with R ...

Recording (Reporting/Responding) Indicator R for a **finite population** $\mathbb{X} = \{X_1, \ldots, X_N\}$:

$R_i = 1$, if X_i is included in a sample of size n, and $R_i = 0$ otherwise

Design Prob: Prob controlled and implemented by human

Ex: Prob sampling, $\Pr(R_i = 1|\mathbb{X}) = n/N$

More: Bootstraps, clinical trials, differential privacy, permutation test, MC.

Divine Prob: Existential Prob by God/Nature or by faith.

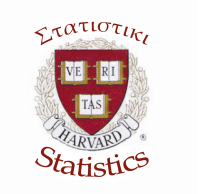
Ex: Real-World Evidence, but we assume R is random, and $\Pr(R_i = 1|\mathbb{X}) = \pi(X_i)$. [Needed for define "missing at Random"]

Device Prob: Probability constructs invoked for analysis.

Ex: $\text{logit}[\pi(x)] = \alpha + \beta x$

More: Prob distributions for expressing belief, prior knowledge, assumptions, idealizations, compromises, desperation.

A Fundamental Identity for Estimation Error (Meng, 2018)



A Fundamental Identity for Estimation Error (Meng, 2018)

- Population $\{X_1, ..., X_N\}$; **Estimand** $\bar{\mu}_N = \frac{\sum_{i=1}^{N} G(X_i)}{N}$;

A Fundamental Identity for Estimation Error (Meng, 2018)

- Population $\{X_1, ..., X_N\}$; **Estimand** $\bar{\mu}_N = \frac{\sum_{i=1}^N G(X_i)}{N}$;
- **Estimator**: sample average

$$\bar{\mu}_n = \frac{\sum_{i=1}^N R_i G(X_i)}{\sum_{i=1}^N R_i} \equiv \frac{\sum_{i=1}^N R_i G(X_i)}{n_R}$$

  where $R_i = 1$ if X_i is recorded, and zero otherwise.
- **Neither X nor R is assumed random**.

A Fundamental Identity for Estimation Error (Meng, 2018)

- Population $\{X_1, ..., X_N\}$; **Estimand** $\bar{\mu}_N = \frac{\sum_{i=1}^N G(X_i)}{N}$;
- **Estimator**: sample average

$$\bar{\mu}_n = \frac{\sum_{i=1}^N R_i G(X_i)}{\sum_{i=1}^N R_i} \equiv \frac{\sum_{i=1}^N R_i G(X_i)}{n_R}$$

where $R_i = 1$ if X_i is recorded, and zero otherwise.

- **Neither X nor R is assumed random**.

Expressing the exact error w.r.t $I \sim U[1, \ldots, N]$ (a divine probability):

$$\bar{\mu}_n - \bar{\mu}_N = \frac{\mathrm{E}_I[R_I G_I]}{\mathrm{E}_I[R_I]} - \mathrm{E}_I[G_I] = \frac{\mathrm{Cov}_I(R_I, G_I)}{\mathrm{E}_I[R_I]}$$

$$= \hat{\rho}_{R,G} \times \sqrt{\frac{N-n}{n}} \times \sigma_G$$

A Fundamental Identity for Estimation Error (Meng, 2018)

- Population $\{X_1, ..., X_N\}$; **Estimand** $\bar{\mu}_N = \frac{\sum_{i=1}^N G(X_i)}{N}$;
- **Estimator**: sample average

$$\bar{\mu}_n = \frac{\sum_{i=1}^N R_i G(X_i)}{\sum_{i=1}^N R_i} \equiv \frac{\sum_{i=1}^N R_i G(X_i)}{n_R}$$

where $R_i = 1$ if X_i is recorded, and zero otherwise.

- **Neither X nor R is assumed random**.

Expressing the exact error w.r.t $I \sim U[1, \ldots, N]$ (a divine probability):

$$\bar{\mu}_n - \bar{\mu}_N = \frac{\mathrm{E}_I[R_I G_I]}{\mathrm{E}_I[R_I]} - \mathrm{E}_I[G_I] = \frac{\mathrm{Cov}_I(R_I, G_I)}{\mathrm{E}_I[R_I]}$$

$$= \hat{\rho}_{R,G} \times \sqrt{\frac{N-n}{n}} \times \sigma_G$$

Why teaching is so important ...

Why teaching is so important ...

Statistics & Probability Letters 18 (1993) 345–348
North-Holland

2 December 1993

On the absolute bias ratio of ratio estimators

Xiao-Li Meng
Department of Statistics, University of Chicago, IL, USA

Received January 1993
Revised March 1993

Abstract: The elegant Hartley–Ross inequality on the absolute bias ratio (ABR ≡ |Bias|/S.E.) of a ordinary ratio estimator is here generalized to that of a separate ratio estimator with stratified sampling. It is shown that, as long as the numerators and denominators used to form strata ratios are unbiased estimators, the absolute bias ratio of a separate ratio estimator will never exceed the square root of the sum of squares of the coefficient of variation of the denominators across strata. This provides, at design stages, a simple bound in practice to assess the limit and magnitude of the bias ratio of any separate ratio estimator that shares the same denominators. Exact expressions for biases of separate ratio estimators are also given.

Keywords: Combined ratio estimator; separate ratio estimator; stratified sampling.

1. Biases of ordinary ratio estimators

In sample surveys, ordinary ratio estimators are typically employed to estimate (i) a population total, Y, (ii) a population mean, $\bar{Y}$, or (iii) a population ratio, Y/X. In all of these cases, the ratio estimator has the form

$$r = \frac{\bar{y}}{\bar{x}} Q, \tag{1.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of variable x and y, respectively, and Q is a known quantity. In cases (i) and (ii), Q is the population total and mean of variable x, X and $\bar{X}$ respectively, and the ratio estimator r of (1.1) is used to increase the precision in estimating Y and $\bar{Y}$ by taking advantage of the positive correlation between y and x and the known values of X and $\bar{X}$ in the population. In case (iii), $Q = 1$, and the population quantities of variable x need not be known. A comprehensive treatment of ratio estimator (1.1) and other variations can be found in Cochran (1977, Chapter 6).

It is well known that in general, r of (1.1) is biased for $R = (\bar{Y}/\bar{X})Q$. However, this bias is typically unimportant because it is negligible compared to the standard error of r. An elementary but elegant proof of this fact was given in Hartley and Ross (1954), who noticed the following simple identity

$$E(r) - \frac{E(\bar{y})}{E(\bar{x})} Q = -\frac{\mathrm{Cov}(r, \bar{x})}{E(\bar{x})}. \tag{1.2}$$

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times$$

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times$$

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_G}_{\textbf{Problem Difficulty}}$$

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_G}_{\textbf{Problem Difficulty}}$$

- Nothing is assumed random, only using divine probability via I.

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_G}_{\textbf{Problem Difficulty}}$$

- Nothing is assumed random, only using divine probability via I.
- Only assumption: sampled X_i are in the target population.

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_G}_{\textbf{Problem Difficulty}}$$

- Nothing is assumed random, only using divine probability via I.
- Only assumption: sampled X_i are in the target population.

Compare to the additive "Variance+Bias" decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\bar{\mu}_n - \mathrm{E}_R(\bar{\mu}_n)}_{\textbf{Sampling Error}} + \underbrace{\mathrm{E}_R(\bar{\mu}_n) - \bar{\mu}_N}_{\textbf{Bias}}$$

A Multiplicative Error Decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\hat{\rho}_{R,G}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_G}_{\textbf{Problem Difficulty}}$$

- Nothing is assumed random, only using divine probability via I.
- Only assumption: sampled X_i are in the target population.

Compare to the additive "Variance+Bias" decomposition

$$\underbrace{\bar{\mu}_n - \bar{\mu}_N}_{\textbf{Exact Error}} = \underbrace{\bar{\mu}_n - \mathrm{E}_R(\bar{\mu}_n)}_{\textbf{Sampling Error}} + \underbrace{\mathrm{E}_R(\bar{\mu}_n) - \bar{\mu}_N}_{\textbf{Bias}}$$

- Must involve prob on R, design, divine, or device.
- It's "phenotype" decomposition, not "genotype."

Why and when can we ignore the population size N?

- Think about tasting soup

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!**

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!**

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!**

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!**

$\Longleftrightarrow$

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!**

$\Longleftrightarrow$

- **But what happens when we fail to stir (well)?**

Probability Sampling miniaturizes the *ddc*

For SRS, because $\mathrm{V}_R(Z) = 1$, where

$$Z = \frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}} S_G} = \hat{\rho}_{R,G}\sqrt{N-1} \quad \Rightarrow \quad \mathrm{V}_R(\hat{\rho}_{R,G}) = \frac{1}{N-1}$$

Hence $\hat{\rho}_{R,G} \in \left(-\frac{3}{\sqrt{N-1}}, \frac{3}{\sqrt{N-1}}\right)$ 99% of the time regardless of G.

Probability Sampling miniaturizes the *ddc*

For SRS, because $\mathrm{V}_R(Z) = 1$, where

$$Z = \frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}} S_G} = \hat{\rho}_{R,G}\sqrt{N-1} \quad \Rightarrow \quad \mathrm{V}_R(\hat{\rho}_{R,G}) = \frac{1}{N-1}$$

Hence $\hat{\rho}_{R,G} \in \left(-\frac{3}{\sqrt{N-1}}, \frac{3}{\sqrt{N-1}}\right)$ 99% of the time regardless of G.

$$\bar{\mu}_n - \bar{\mu}_N = \underbrace{\hat{\rho}_{R,G}}_{1/\sqrt{N}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\sqrt{N}} \times \sigma_G$$

Probability Sampling miniaturizes the *ddc*

For SRS, because $\mathrm{V}_R(Z) = 1$, where

$$Z = \frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}} S_G} = \hat{\rho}_{R,G}\sqrt{N-1} \quad \Rightarrow \quad \mathrm{V}_R(\hat{\rho}_{R,G}) = \frac{1}{N-1}$$

Hence $\hat{\rho}_{R,G} \in \left(-\frac{3}{\sqrt{N-1}}, \frac{3}{\sqrt{N-1}}\right)$ 99% of the time regardless of G.

$$\bar{\mu}_n - \bar{\mu}_N = \underbrace{\hat{\rho}_{R,G}}_{1/\sqrt{N}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\sqrt{N}} \times \sigma_G$$

- This cancellation is **THE reason** that we can ignore N **for any** G.

Data Defect Correlation (ddc) and Data Defect Index (ddi)

ddc: $\hat{\rho}_{R,G}$

$$\frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}S_G}} = \hat{\rho}_{R,G}\sqrt{N-1}$$

ddi: $D_I = \mathrm{E}_R(\hat{\rho}^2_{R,G})$

$$\mathrm{Deff} \equiv \frac{\mathrm{MSE}(\bar{\mu}_n)}{\mathrm{SRS\ MSE}} = D_I(N-1)$$

Data Defect Correlation (ddc) and Data Defect Index (ddi)

ddc: $\hat{\rho}_{R,G}$

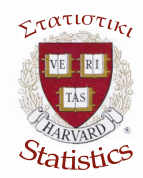
$$\frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}S_G}} = \hat{\rho}_{R,G}\sqrt{N-1}$$

ddi: $D_I = \mathrm{E}_R(\hat{\rho}^2_{R,G})$

$$\mathrm{Deff} \equiv \frac{\mathrm{MSE}(\bar{\mu}_n)}{\text{SRS MSE}} = D_I(N-1)$$

ddi $= \frac{\mathrm{Deff}}{N-1} =$ Design effect per subject in the population

- For SRS: $D_I = (N-1)^{-1}$



Data Defect Correlation (ddc) and Data Defect Index (ddi)

ddc: $\hat{\rho}_{R,G}$

$$\frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}S_G}} = \hat{\rho}_{R,G}\sqrt{N-1}$$

ddi: $D_I = \mathrm{E}_R(\hat{\rho}^2_{R,G})$

$$\text{Deff} \equiv \frac{\text{MSE}(\bar{\mu}_n)}{\text{SRS MSE}} = D_I(N-1)$$

ddi $= \frac{\text{Deff}}{N-1} =$ Design effect per subject in the population

- For SRS: $D_I = (N-1)^{-1}$
- Probability sample $\Longrightarrow$ $D_I \propto N^{-1}$

Data Defect Correlation (ddc) and Data Defect Index (ddi)

ddc: $\hat{\rho}_{R,G}$

$$\frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n} S_G}} = \hat{\rho}_{R,G} \sqrt{N-1}$$

ddi: $D_I = \mathrm{E}_R(\hat{\rho}^2_{R,G})$

$$\mathrm{Deff} \equiv \frac{\mathrm{MSE}(\bar{\mu}_n)}{\text{SRS MSE}} = D_I(N-1)$$

ddi $= \frac{\mathrm{Deff}}{N-1} =$ Design effect per subject in the population

- For SRS: $D_I = (N-1)^{-1}$
- Probability sample $\Longrightarrow$ $D_I \propto N^{-1}$
- $MSE(\bar{\mu}_n) \propto n^{-1}$ $\Longleftrightarrow$ $D_I \propto N^{-1}$

Data Defect Correlation (ddc) and Data Defect Index (ddi)

ddc: $\hat{\rho}_{R,G}$

$$\frac{\bar{\mu}_n - \bar{\mu}_N}{\sqrt{\frac{1-f}{n}S_G}} = \hat{\rho}_{R,G}\sqrt{N-1}$$

ddi: $D_I = \mathrm{E}_R(\hat{\rho}^2_{R,G})$

$$\mathrm{Deff} \equiv \frac{\mathrm{MSE}(\bar{\mu}_n)}{\text{SRS MSE}} = D_I(N-1)$$

ddi $= \frac{\mathrm{Deff}}{N-1} =$ **Design effect per subject in the population**

- For SRS: $D_I = (N-1)^{-1}$
- Probability sample $\Longrightarrow$ $D_I \propto N^{-1}$
- $MSE(\bar{\mu}_n) \propto n^{-1}$ $\Longleftrightarrow$ $D_I \propto N^{-1}$

Deep Trouble

- when D_I does not vanish with N^{-1};
- or equivalently when $\hat{\rho}_{G,R}$ does not vanish with $N^{-1/2}$ ...

How Small is Our "Big Data"?

How Small is Our "Big Data"?

The *Effective Sample Size* of a "Big Data" in terms of SRS size

$$n_{\text{eff}} = \frac{n}{1 + (1-f)[(N-1)\mathsf{E}_R[\hat{\rho}^2_{R,G}] - 1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}$$

How Small is Our "Big Data"?

The *Effective Sample Size* of a "Big Data" in terms of SRS size

$$n_{\text{eff}} = \frac{n}{1 + (1-f)[(N-1)\mathsf{E}_R[\hat{\rho}^2_{R,G}] - 1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}$$

Why do we need random testing? (Walter Dempsey, Twitter)

- NY State: $N \approx 20$M;

How Small is Our "Big Data"?

The *Effective Sample Size* of a "Big Data" in terms of SRS size

$$n_{\text{eff}} = \frac{n}{1 + (1-f)[(N-1)\mathsf{E}_R[\hat{\rho}^2_{R,G}] - 1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}$$

Why do we need random testing? (Walter Dempsey, Twitter)

- NY State: $N \approx 20\text{M}$;
- Suppose we conduct $n = 10,000$ COVID test: $f = 1/2000$

How Small is Our "Big Data"?

The *Effective Sample Size* of a "Big Data" in terms of SRS size

$$n_{\text{eff}} = \frac{n}{1 + (1-f)[(N-1)\mathsf{E}_R[\hat{\rho}^2_{R,G}] - 1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}$$

Why do we need random testing? (Walter Dempsey, Twitter)

- NY State: $N \approx 20\text{M}$;
- Suppose we conduct $n = 10,000$ COVID test: $f = 1/2000$
- Suppose the selective testing resulted in $\hat{\rho} = 0.005$;

$$n_{\text{eff}} = \frac{0.0005}{0.9995} \times \frac{1}{0.005^2} \approx 20$$

How Small is Our "Big Data"?

The *Effective Sample Size* of a "Big Data" in terms of SRS size

$$n_{\text{eff}} = \frac{n}{1 + (1-f)[(N-1)\mathsf{E}_R[\hat{\rho}^2_{R,G}] - 1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}$$

Why do we need random testing? (Walter Dempsey, Twitter)

- NY State: $N \approx 20\text{M}$;
- Suppose we conduct $n = 10,000$ COVID test: $f = 1/2000$
- Suppose the selective testing resulted in $\hat{\rho} = 0.005$;

$$n_{\text{eff}} = \frac{0.0005}{0.9995} \times \frac{1}{0.005^2} \approx 20$$

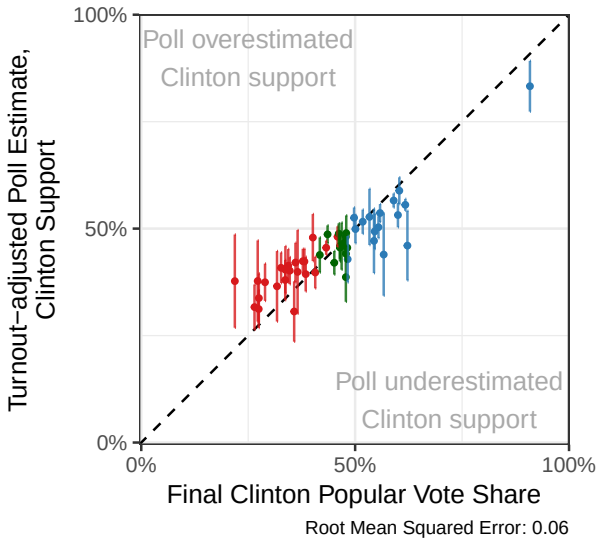
- Hence $\hat{\rho} = 0.005$ implies a 99.80% loss of sample size!

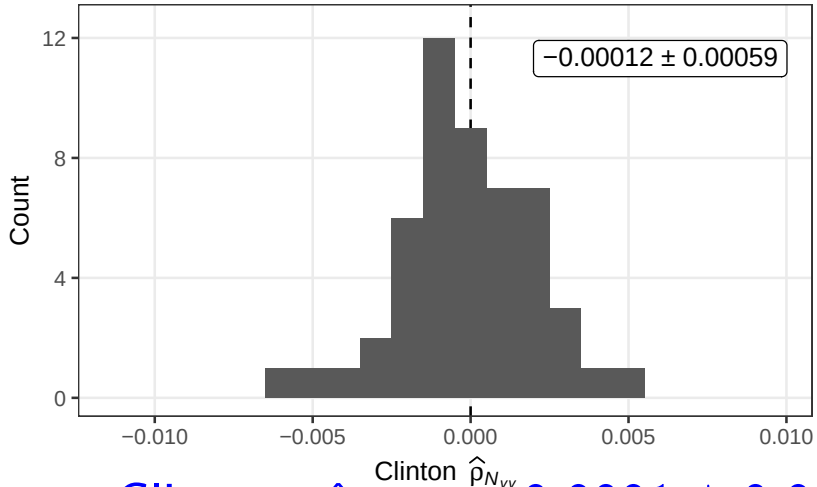
Where Did $\hat{\rho} = 0.005$ Come From? (Meng, 2018)

Cooperative Congressional Election Study by Ansolabehere, Schaffner, Luks, Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki

Where Did $\hat{\rho} = 0.005$ Come From? (Meng, 2018)

Cooperative Congressional Election Study by Ansolabehere, Schaffner, Luks, Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki

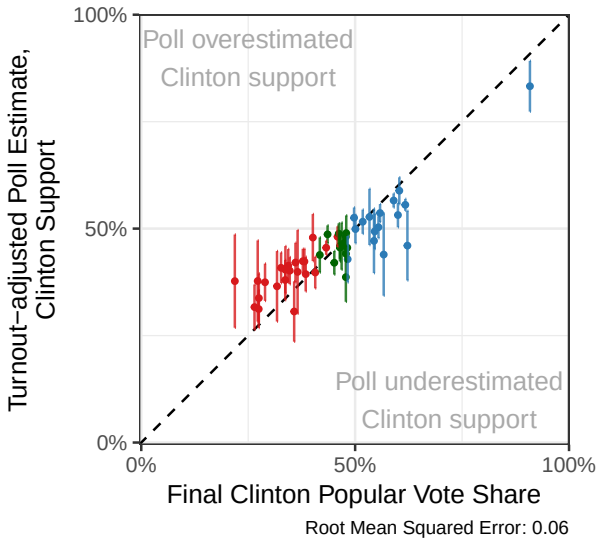


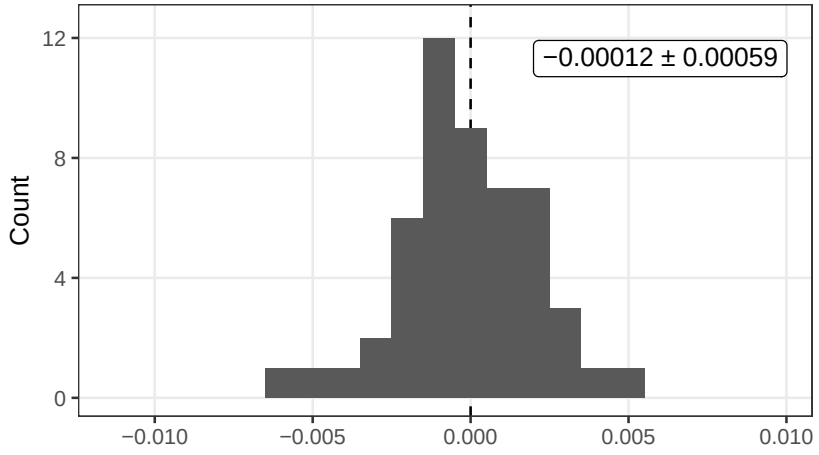


Clinton: $\hat{\rho}_{vvt} \approx -0.0001 \pm 0.0006$

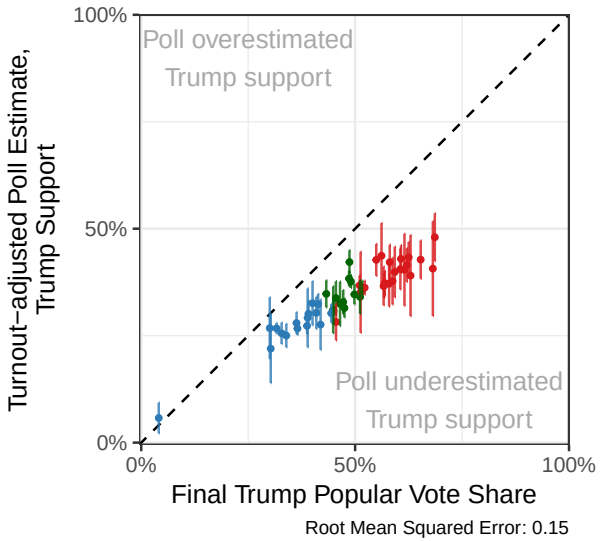
Where Did $\hat{\rho} = 0.005$ Come From? (Meng, 2018)

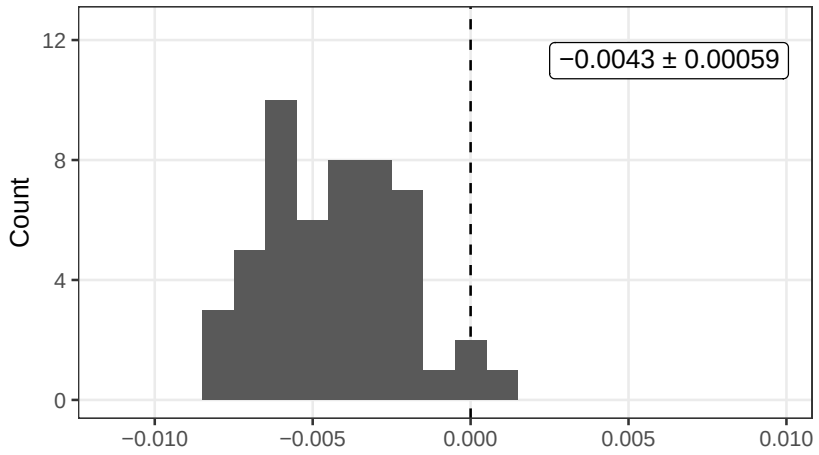
Cooperative Congressional Election Study by Ansolabehere, Schaffner, Luks, Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki





Clinton: $\hat{\rho}_{vvt} \approx -0.0001 \pm 0.0006$

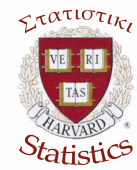




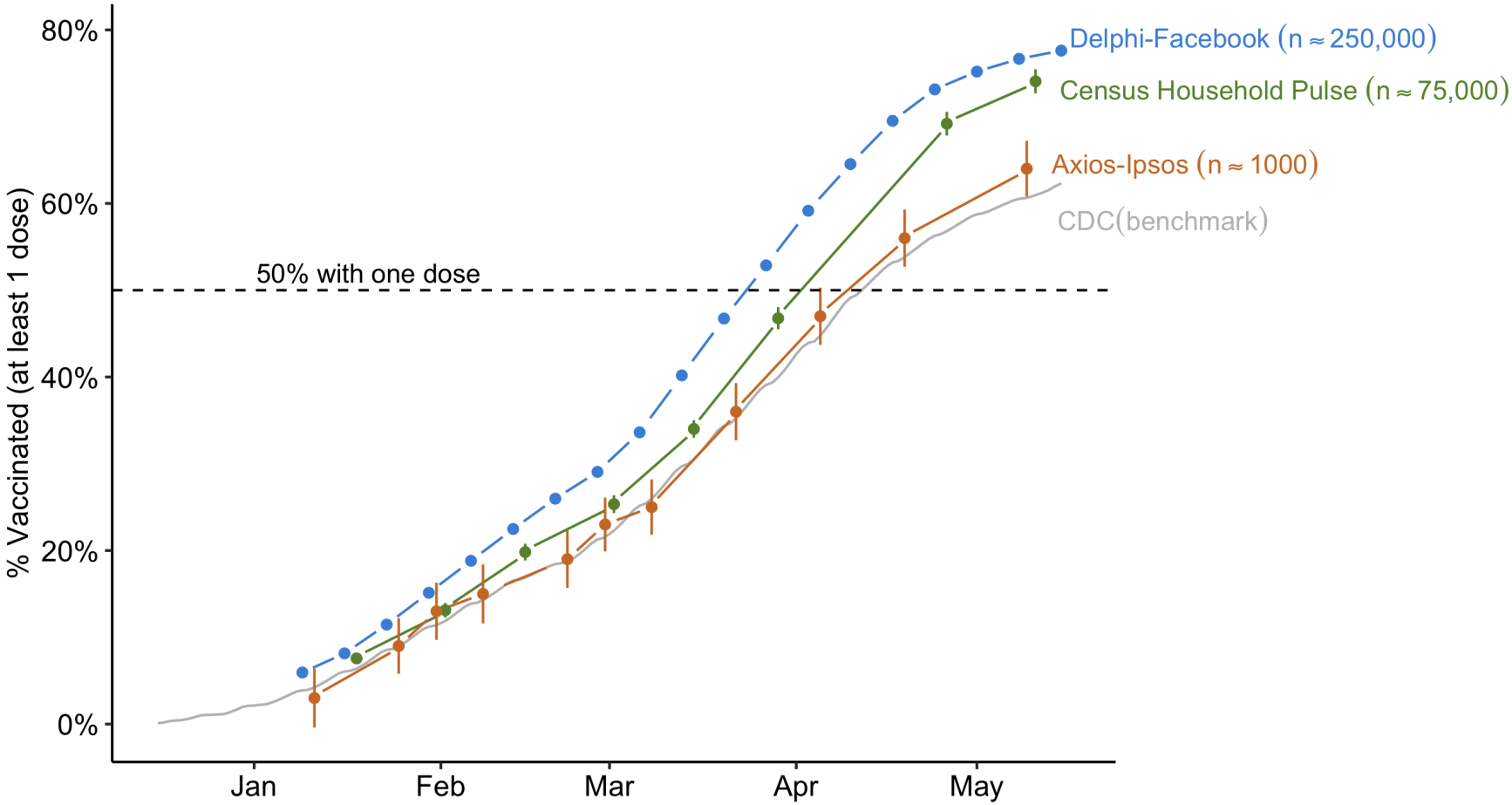
Trump: $\hat{\rho}_{vvt} \approx -0.0043 \pm 0.0006$

Studies of COVID-19 vaccine uptake vary in design

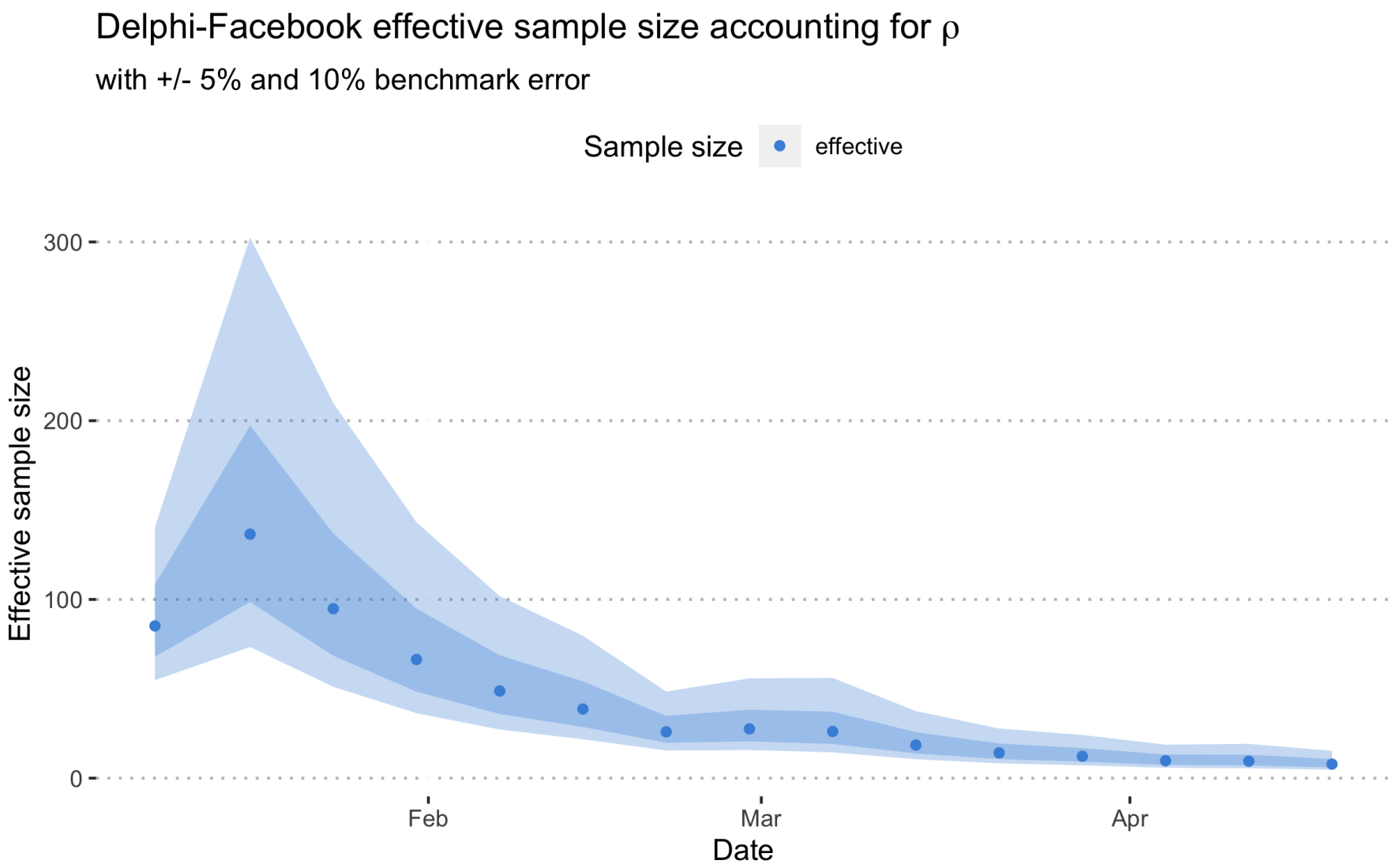
	Axios-Ipsos Coronavirus Index	**Census Household Pulse**	**Facebook-Delphi COVID Symptom Survey**
Mode	online	online	online
Sampling frame	Ipsos KnowledgePanel	Census Master Address File	Facebook Active User Base
n	1,000/wave	65,000/wave	250,000/wave
Question Wording	"Do you personally know anyone who has already received the COVID-19 vaccine?" Answers include "Yes, I have received the vaccine"	"Have you received a COVID-19 vaccine?"	"Have you had a COVID-19 vaccination?"
Target Population	2019 CPS March Supplement, US 18+	2018 ACS, 1-year est., US 18+	2018 CPS March Supplement, US 18+
Weighting Variables	gender x age, race, education, Census region, metropolitan status, household income, partisanship	metropolitan statistical area (MSA), state x education x gender x age, state x Hispanic ethnicity x gender x age	state x age x gender and "proprietary covariates"



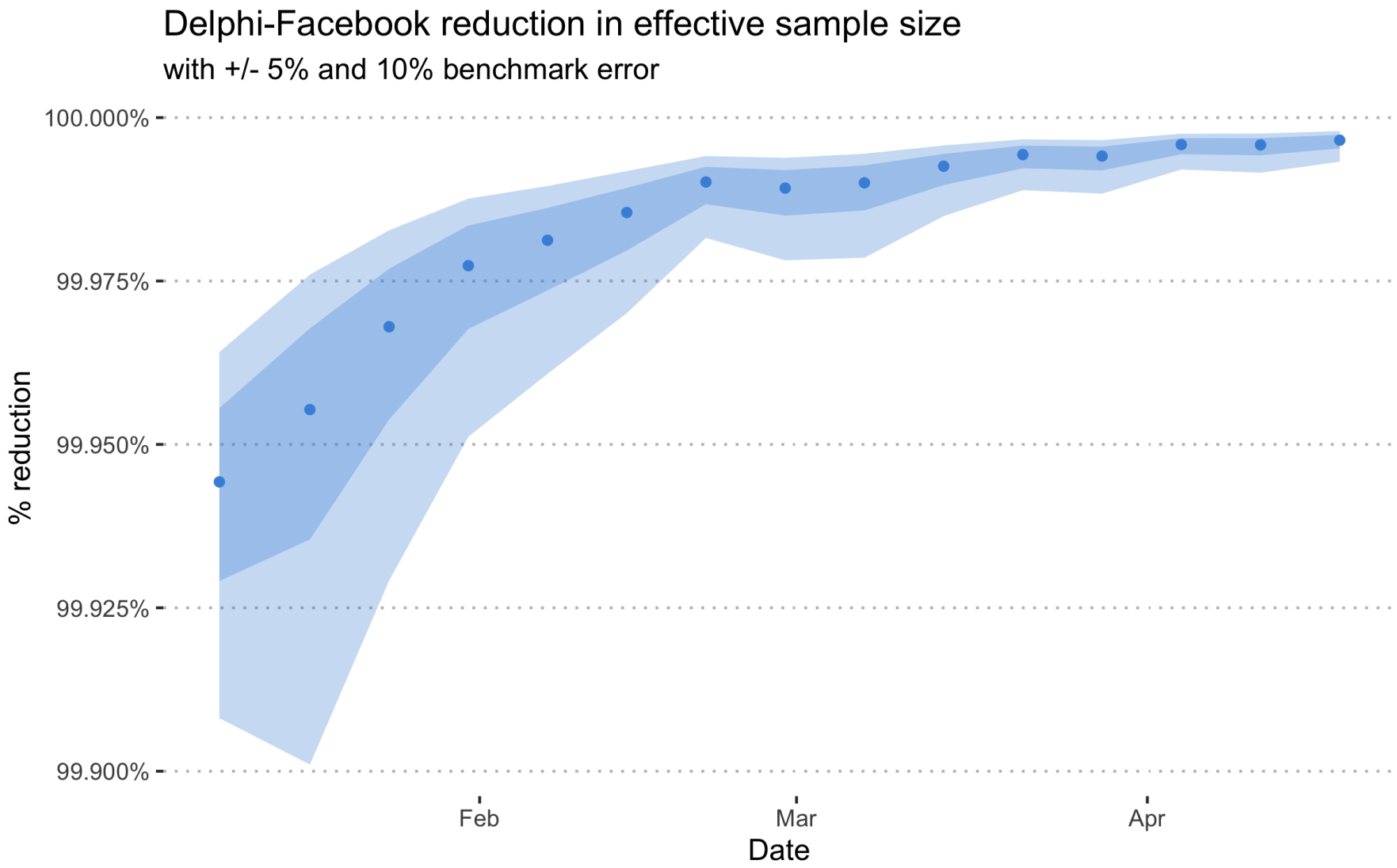
Big Data Paradox: The Bigger the Data, The Surer We Fool Ourselves



Estimated Effective Sample Sizes (dropping three 0s!)



Dramatic Reduction in Effective Sample Size ($> 99.9\%$)



Composition of Survey Respondents

	Composition of U.S. Adults							Survey Estimates		
	Axios-Ipsos		Household Pulse		Delphi-Facebook		ACS	Household Pulse		
Education	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
High School	35%	39%	14%	39%	19%	21%	39%	39%	40%	21%
Some College	29	30	32	30	36	36	30	44	38	18
4-Year College	19	17	29	17	25	25	19	54	36	10
Post-Graduate	17	14	26	13	20	18	11	67	26	7

	Composition of U.S. Adults							Survey Estimates		
	Axios-Ipsos		Household Pulse		Delphi-Facebook		ACS	Household Pulse		
Race/Ethnicity	Raw	Weighted	Raw	Weighted	Raw	Weighted	Benchmark	Vax	Will	Hes
White	71%	63%	75%	62%	74%	68%	60%	50%	33%	17%
Black	10	12	7	11	6	6	12	42	39	19
Hispanic	11	16	10	17	11	16	16	38	48	14
Asian			5	5	2	3	6	51	43	5

"Weight, Weight, Don't tell me ..."

There are always those "messy" weights ...

$$\bar{\mu}_w = \frac{\sum_{i=1}^{N} R_i w_i G(X_i)}{\sum_{i=1}^{N} R_i w_i}$$

"Weight, Weight, Don't tell me ..."

There are always those "messy" weights ...

$$\bar{\mu}_w = \frac{\sum_{i=1}^{N} R_i w_i G(X_i)}{\sum_{i=1}^{N} R_i w_i}$$

Let CV_w be the coefficient of variation of W_I given $R_I = 1$

$$\bar{\mu}_w - \bar{\mu}_N = \hat{\rho}_{Rw,G} \times \sqrt{\frac{1 - f_w}{f_w}} \times \sigma_G$$

where

$$f_w = \frac{n_w}{N}, \qquad n_w = \frac{n}{1 + CV_w^2} \quad \text{(Kish, 1965)}$$

"Weight, Weight, Don't tell me ..."

There are always those "messy" weights ...

$$\bar{\mu}_w = \frac{\sum_{i=1}^{N} R_i w_i G(X_i)}{\sum_{i=1}^{N} R_i w_i}$$

Let CV_w be the coefficient of variation of W_I given $R_I = 1$

$$\bar{\mu}_w - \bar{\mu}_N = \hat{\rho}_{Rw,G} \times \sqrt{\frac{1 - f_w}{f_w}} \times \sigma_G$$

where

$$f_w = \frac{n_w}{N}, \qquad n_w = \frac{n}{1 + CV_w^2} \quad \text{(Kish, 1965)}$$

- Seeking w to make $\hat{\rho}_{Rw,G} < \hat{\rho}_{R,G}$ and compensate for $n_w < n$.

Estimation Methods for Non-probability Samples

- A non-probability sample $\{(y_i, X_i), i \in S\}, \quad S = \{i : R_i = 1\}$
- An auxiliary probability sample $\{X_i, i \in S^*\}, \quad S^* = \{i : R_i^* = 1\}$
- Key Assumption: $y_i \perp R_i | X_i$ (Missing at Random)
- A device model: $p(y, R|x) = p(y|x)p(R|x)$

Estimation Methods for Non-probability Samples

- A non-probability sample $\{(y_i, X_i), i \in S\}, \quad S = \{i : R_i = 1\}$
- An auxiliary probability sample $\{X_i, i \in S^*\}, \quad S^* = \{i : R_i^* = 1\}$
- Key Assumption: $y_i \perp R_i | X_i$ (Missing at Random)
- A device model: $p(y, R|x) = p(y|x)p(R|x)$

Quasi-randomization: Estimate $\pi(x) = \Pr_p(R = 1|x)$

$$\hat{\mu} = \frac{\sum_{i=1}^N R_i w_i y_i}{\sum_{i=1}^N R_i w_i}, \qquad w_i \propto \hat{\pi}^{-1}(X_i)$$

Estimation Methods for Non-probability Samples

- A non-probability sample $\{(y_i, X_i), i \in S\}, \quad S = \{i : R_i = 1\}$
- An auxiliary probability sample $\{X_i, i \in S^*\}, \quad S^* = \{i : R_i^* = 1\}$
- Key Assumption: $y_i \perp R_i | X_i$ (Missing at Random)
- A device model: $p(y, R|x) = p(y|x)p(R|x)$

Quasi-randomization: Estimate $\pi(x) = \Pr_p(R = 1|x)$

$$\hat{\mu} = \frac{\sum_{i=1}^N R_i w_i y_i}{\sum_{i=1}^N R_i w_i}, \qquad w_i \propto \hat{\pi}^{-1}(X_i)$$

Super-population/model assisted: also fit $y = m(x)$

$$\hat{\mu}_+ = \frac{\sum_{i=1}^N R_i w_i (y_i - \hat{m}(X_i))}{\sum_{i=1}^N R_i w_i} + \frac{\sum_{i=1}^N R_i^* \hat{m}(X_i)}{\sum_{i=1}^N R_i^*},$$

Why Do They Work?

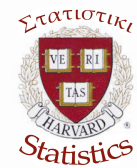
$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Why Do They Work?

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Under a divine model for $(R, Y|X)$ and sampling model for R^*

$$\mathrm{E}[\hat{\mu}_+] - \mu \sim \mathrm{E}_x\left\{\mathrm{Cov}_I[\pi_I w_I, \delta_I]\right\}, \quad \text{where } \delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I)$$



Why Do They Work?

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Under a divine model for $(R, Y|X)$ and sampling model for R^*

$$\mathrm{E}[\hat{\mu}_+] - \mu \sim \mathrm{E}_x\left\{\mathrm{Cov}_I[\pi_I w_I, \delta_I]\right\}, \quad \text{where } \delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I)$$

- **Quasi-randomization**: making $\pi_I w_I \propto 1$, $\qquad (Q)$

Why Do They Work?

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Under a divine model for $(R, Y|X)$ and sampling model for R^*

$$\mathrm{E}[\hat{\mu}_+] - \mu \sim \mathrm{E}_x\left\{\mathrm{Cov}_I[\pi_I w_I, \delta_I]\right\}, \quad \text{where } \delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I)$$

- **Quasi-randomization**: making $\pi_I w_I \propto 1$, $\quad (Q)$
- **Super-population:** making $\delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I) = 0$, $\quad (S)$

Why Do They Work?

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Under a divine model for $(R, Y|X)$ and sampling model for R^*

$$\mathrm{E}[\hat{\mu}_+] - \mu \sim \mathrm{E}_x\left\{\mathrm{Cov}_I[\pi_I w_I, \delta_I]\right\}, \quad \text{where } \delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I)$$

- **Quasi-randomization**: making $\pi_I w_I \propto 1$, $\quad (Q)$
- **Super-population:** making $\delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I) = 0$, $\quad (S)$
- **Doubly robust**: either (Q) or (S) makes $\mathrm{Cov}(\pi_I w_I, \delta_I) = 0$, but we don't need to know which one.

Why Do They Work?

$$\hat{\mu}_+ - \bar{y}_N = \frac{\mathrm{Cov}_I(R_I w_I, y_I - \hat{m}(X_I))}{\mathrm{E}_I(R_I w_I)} + \frac{\mathrm{Cov}_I(R_I^*, \hat{m}(X_I))}{\mathrm{E}_I(R_I^*)},$$

Under a divine model for $(R, Y|X)$ and sampling model for R^*

$$\mathrm{E}[\hat{\mu}_+] - \mu \sim \mathrm{E}_x\left\{\mathrm{Cov}_I[\pi_I w_I, \delta_I]\right\}, \quad \text{where } \delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I)$$

- **Quasi-randomization**: making $\pi_I w_I \propto 1$, $\quad (Q)$
- **Super-population:** making $\delta_I \equiv \mathrm{E}(y_I|X_I) - \hat{m}(X_I) = 0$, $\quad (S)$
- **Doubly robust**: either (Q) or (S) makes $\mathrm{Cov}(\pi_I w_I, \delta_I) = 0$, but we don't need to know which one.
- **"Double+ robustness"**: the validity holds if and only if

$$\mathrm{E}_x\left\{\mathrm{Cov}_I\left[\pi_I w_I, \delta_I\right]\right\} = 0$$

Counterbalance Sub-sampling (CBS)

- If $\Pr(R_i = 1|X, Y) = \pi_i$, sub-sample with $\Pr(S_i = 1|R_i = 1) \propto \pi_i^{-1}$

Counterbalance Sub-sampling (CBS)

- If $\Pr(R_i = 1|X, Y) = \pi_i$, sub-sample with $\Pr(S_i = 1|R_i = 1) \propto \pi_i^{-1}$
- Hence $\Pr(S_i R_i = 1) \propto 1$, creating an equal-probability sample

Counterbalance Sub-sampling (CBS)

- If $\Pr(R_i = 1|X, Y) = \pi_i$, sub-sample with $\Pr(S_i = 1|R_i = 1) \propto \pi_i^{-1}$
- Hence $\Pr(S_i R_i = 1) \propto 1$, creating an equal-probability sample

The *Effective Sample Size* of "Big Data" in terms of SRS size

$$n_B \approx \frac{f_B}{1 - f_B} \frac{1}{\hat{\rho}_B^2},$$

where $f_B = n/N$ is the *relative size* of sample B.

Counterbalance Sub-sampling (CBS)

- If $\Pr(R_i = 1|X, Y) = \pi_i$, sub-sample with $\Pr(S_i = 1|R_i = 1) \propto \pi_i^{-1}$
- Hence $\Pr(S_i R_i = 1) \propto 1$, creating an equal-probability sample

The *Effective Sample Size* of "Big Data" in terms of SRS size

$$n_B \approx \frac{f_B}{1 - f_B} \frac{1}{\hat{\rho}_B^2},$$

where $f_B = n/N$ is the *relative size* of sample B.

Trading quantity for quality: create an unweighted sub-sample D such that

$$\frac{f_B}{1 - f_B} \frac{1}{\hat{\rho}_B^2} < \frac{f_D}{1 - f_D} \frac{1}{\hat{\rho}_D^2},$$

where $f_D = f_B f_S$, and f_S is the sub-sampling rate.

Estimand $p = P(Y = 1)$; estimable $p^* = P(Y = 1|R = 1)$

Estimand $p = P(Y = 1)$; estimable $p^* = P(Y = 1|R = 1)$

- Reporting rates:

$$r_y = \Pr(R = 1|Y = y); \quad r = \frac{r_1}{r_0}$$

- Sub-sampling rates:

$$s_y = \Pr(S = 1|R = 1, Y = y); \qquad s = \frac{s_1}{s_0}$$

Estimand $p = P(Y = 1)$; estimable $p^* = P(Y = 1|R = 1)$

- Reporting rates:

$$r_y = \Pr(R = 1|Y = y); \quad r = \frac{r_1}{r_0}$$

- Sub-sampling rates:

$$; \qquad s_y = \Pr(S = 1|R = 1, Y = y); \qquad s = \frac{s_1}{s_0}$$

Counterbalancing: $(r - 1)(s - 1) < 0$

(i) If $\delta = r - 1 > 0$, then take any s

$$\frac{[1 - (1 - p^*)\delta]_+}{1 + (1 + p^*)\delta} \leq s < 1 \tag{1}$$

(ii) If $\delta = r - 1 < 0$, then take any s

$$1 < s \leq \frac{1 - (1 - p^*)\delta}{[1 + (1 + p^*)\delta]_+} \tag{2}$$

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

We do not know r, but

- Suppose a previous survey had $r = 1.5$

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

We do not know r, but

- Suppose a previous survey had $r = 1.5$
- We might feel comfortable to assume that the current $r \in (1.2, 1.8)$

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

We do not know r, but

- Suppose a previous survey had $r = 1.5$
- We might feel comfortable to assume that the current $r \in (1.2, 1.8)$
- Suppose we observe $p^* = 0.6$

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

We do not know r, but

- Suppose a previous survey had $r = 1.5$
- We might feel comfortable to assume that the current $r \in (1.2, 1.8)$
- Suppose we observe $p^* = 0.6$
- Then the max of the lower bound is 0.7

An miniature illustration

Counterbalancing: $(r-1)(s-1) < 0$

When $\delta = r - 1 > 0$, take any s

$$\frac{[1-(1-p^*)\delta]_+}{1+(1+p^*)\delta} \leq s < 1$$

We do not know r, but

- Suppose a previous survey had $r = 1.5$
- We might feel comfortable to assume that the current $r \in (1.2, 1.8)$
- Suppose we observe $p^* = 0.6$
- Then the max of the lower bound is 0.7
- If $r = 1.5$, then $s = 1/1.5 = 0.67$ will be optimal, but any $s \in [0.7, 1)$ will lead to smaller MSE compared to not using CBS

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off
- Data cleanness and data relevance trade-off

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off
- Data cleanness and data relevance trade-off
- Data utility and data privacy trade-off

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off
- Data cleanness and data relevance trade-off
- Data utility and data privacy trade-off
- ...

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off
- Data cleanness and data relevance trade-off
- Data utility and data privacy trade-off
- ...

For those who just woke up ...

There is no free lunch

- Data quality and data quantity trade-off
- Data cleanness and data relevance trade-off
- Data utility and data privacy trade-off
- ...

But let's not overpay either

- Avoid add more noise than necessary
- Avoid weighted data when an unweighted sub-sample can provide similar statistical information
- ...