How Small Are Our Big Data:

Turning the 2016 Surprise into a 2020 Vision

Xiao-Li Meng
Department of Statistics, Harvard University

How Small Are Our Big Data:
Turning the 2016 Surprise into a 2020 Vision

Xiao-Li Meng
Department of Statistics, Harvard University

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

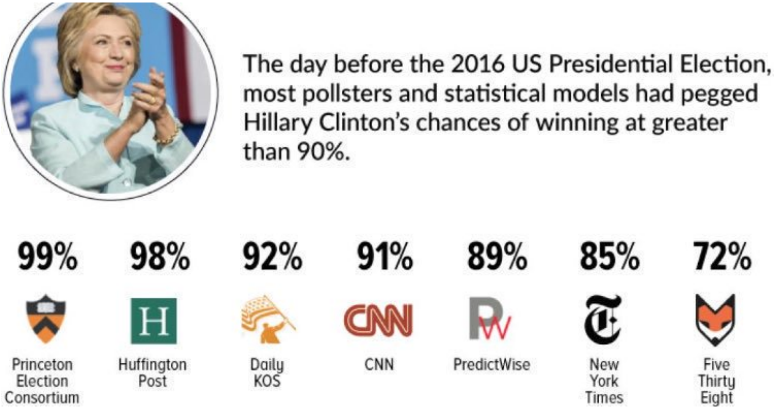
How Small Are Our Big Data:
Turning the 2016 Surprise into a 2020 Vision

Xiao-Li Meng
Department of Statistics, Harvard University

- Meng (2018) **Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and The 2016 US Election**.
  The Annals of Applied Statistics Vol 2: 685-726
- Many thanks to **Stephen Ansolabehere and Shiro Kuriwaki** for the CCES (**Cooperative Congressional Election Study**) data and analysis on 2016 US election.

A Painful Reminder ... (from internet)

Xiao-Li Meng
Department of Statistics, Harvard University



A Chinese survey has size n; a US survey has size m. What should the ratio n/m be for the two surveys to have similar statistical accuracy?

A Chinese survey has size n; a US survey has size m. What should the ratio n/m be for the two surveys to have similar statistical accuracy?

Xiao-Li Meng
Department of Statistics, Harvard University

- Think about tasting soup ...

A Chinese survey has size n; a US survey has size m. What should the ratio n/m be for the two surveys to have similar statistical accuracy?

- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

A Chinese survey has size n; a US survey has size m. What should the ratio n/m be for the two surveys to have similar statistical accuracy?

Xiao-Li Meng
Department of Statistics, Harvard University

- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

Xiao-Li Meng
Department of Statistics, Harvard University

A Chinese survey has size n; a US survey has size m. What should the ratio n/m be for the two surveys to have similar statistical accuracy?

- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**


$\Longleftrightarrow$

2016 US Presidential Election

- n: number of respondents to an election survey

2016 US Presidential Election

- n: number of respondents to an election survey
- N: number of (actual) voters in US

2016 US Presidential Election

- n: number of respondents to an election survey
- N: number of (actual) voters in US
- $X_j = 1$: plan to vote for Trump; $X_j = 0$ otherwise

2016 US Presidential Election

- n: number of respondents to an election survey
- N: number of (actual) voters in US
- $X_j = 1$: plan to vote for Trump; $X_j = 0$ otherwise
- $R_j = 1$: report (honestly) voting plan; $R_j = 0$ otherwise

2016 US Presidential Election

Xiao-Li Meng
Department of Statistics, Harvard University

- n: number of respondents to an election survey
- N: number of (actual) voters in US
- $X_j = 1$: plan to vote for Trump; $X_j = 0$ otherwise
- $R_j = 1$: report (honestly) voting plan; $R_j = 0$ otherwise

2016 US Presidential Election

- n: number of respondents to an election survey
- N: number of (actual) voters in US
- $X_j = 1$: plan to vote for Trump; $X_j = 0$ otherwise
- $R_j = 1$: report (honestly) voting plan; $R_j = 0$ otherwise

Estimatinng Trump's share: $\mu_N = \text{Ave}(X_j)$ by sample average:

$$\hat{\mu}_n = \frac{R_1X_1 + \ldots + R_NX_N}{R_1 + \ldots + R_N} = \frac{\text{Ave}(R_jX_j)}{\text{Ave}(R_j)}$$

Xiao-Li Meng
Department of Statistics, Harvard University

2016 US Presidential Election

- n: number of respondents to an election survey
- N: number of (actual) voters in US
- $X_j = 1$: plan to vote for Trump; $X_j = 0$ otherwise
- $R_j = 1$: report (honestly) voting plan; $R_j = 0$ otherwise

Estimatinng Trump's share: $\mu_N = \text{Ave}(X_j)$ by sample average:

$$\hat{\mu}_n = \frac{R_1 X_1 + \ldots + R_N X_N}{R_1 + \ldots + R_N} = \frac{\text{Ave}(R_j X_j)}{\text{Ave}(R_j)}$$

Actual estimation error

$$\begin{aligned}\hat{\mu}_n - \mu_N &= \frac{\text{Ave}(R_j X_j)}{\text{Ave}(R_j)} - \text{Ave}(X_j) \\ &= \left[\frac{\text{Ave}(R_j X_j) - \text{Ave}(R_j)\text{Ave}(X_j)}{\sigma_R \sigma_X}\right] \times \frac{\sigma_R}{\text{Ave}(R_j)} \times \sigma_X\end{aligned}$$

Data quality, quantity, and uncertainty

Because $\sigma_R^2 = f(1-f)$, $f = \text{Ave}\{R_j\} = \frac{n}{N}$, we have

$$\text{Error} = \underbrace{\hat{\rho}_{R,X}}_{\textbf{Data Quality}} \times$$

Because $\sigma_R^2 = f(1-f)$, $f = \text{Ave}\{R_j\} = \frac{n}{N}$, we have

$$\text{Error} = \underbrace{\hat{\rho}_{R,X}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times$$

Data quality, quantity, and uncertainty

Because $\sigma_R^2 = f(1-f)$, $f = \text{Ave}\{R_j\} = \frac{n}{N}$, we have

$$\text{Error} = \underbrace{\hat{\rho}_{R,X}}_{\textbf{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\textbf{Data Quantity}} \times \underbrace{\sigma_X}_{\textbf{Problem Difficulty}}$$

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mu}_n) = \mathsf{E}_R(\hat{\rho}^2) \times \frac{N-n}{n} \times \sigma_X^2$$

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mu}_n) = \mathsf{E}_R(\hat{\rho}^2) \times \frac{N-n}{n} \times \sigma_X^2$$

Data Defect Index (d.d.i): $D_I = \mathsf{E}_R(\hat{\rho}^2)$

Xiao-Li Meng
Department of Statistics, Harvard University

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mu}_n) = \mathsf{E}_R(\hat{\rho}^2) \times \frac{N-n}{n} \times \sigma_x^2$$

Data Defect Index (d.d.i): $D_I = \mathsf{E}_R(\hat{\rho}^2)$

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

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mu}_n) = \mathsf{E}_R(\hat{\rho}^2) \times \frac{N-n}{n} \times \sigma_X^2$$

Data Defect Index (d.d.i): $D_I = \mathsf{E}_R(\hat{\rho}^2)$

- For Simple Random Sample (SRS): $D_I = (N-1)^{-1}$
- For probabilistic samples in general: $D_I \propto N^{-1}$

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

$$\mathrm{MSE}(\hat{\mu}_n) = \mathsf{E}_R(\hat{\rho}^2) \times \frac{N-n}{n} \times \sigma_x^2$$

Data Defect Index (d.d.i): $D_I = \mathsf{E}_R(\hat{\rho}^2)$

- For Simple Random Sample (SRS): $D_I = (N-1)^{-1}$
- For probabilistic samples in general: $D_I \propto N^{-1}$
- Deep trouble when D_I does not vanish with N^{-1};
- or equivalently when $\hat{\rho}$ does not vanish with $N^{-1/2}$ ...

A Law of Large Populations (LLP)

Xiao-Li Meng
Department of Statistics, Harvard University

If $\rho = \mathrm{E}_R(\hat{\rho}) \neq 0$, then on average, the relative error $\uparrow \sqrt{N}$:

$$\frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

A Law of Large Populations (LLP)

If $\rho = \mathrm{E}_R(\hat{\rho}) \neq 0$, then on average, the relative error $\uparrow \sqrt{N}$:

$$\frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

The (lack-of) design effect (Deff)

$$\text{Deff} = \frac{\text{MSE}}{\text{Benchmark SRS MSE}} = (N-1)D_I$$

A Law of Large Populations (LLP)

If $\rho = \mathrm{E}_R(\hat{\rho}) \neq 0$, then on average, the relative error $\uparrow \sqrt{N}$:

$$\frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

The (lack-of) design effect (Deff)

$$\text{Deff} = \frac{\text{MSE}}{\text{Benchmark SRS MSE}} = (N-1)D_I$$

Paradigm shift for "Big Data":

$$\mathit{From} \quad \underbrace{\frac{\sigma}{\sqrt{n}}}_{\mathit{random\ \ error}} \quad \mathit{to} \quad \underbrace{\hat{\rho}\sqrt{N}}_{\mathit{relative\ \ systemtic\ \ bias}}$$

Effective Sample Size

Xiao-Li Meng
Department of Statistics, Harvard University

Effective Sample Size

The *Effective Sample Size* n_{eff} of a "Big Data" set

Equate its MSE to that from a SRS with size n_{eff}:

$$D_I \left[\frac{N-n}{n}\right]\sigma^2 = \frac{1}{N-1}\left[\frac{N-n_{\text{eff}}}{n_{\text{eff}}}\right]\sigma^2$$

Effective Sample Size

The *Effective Sample Size* $n_{\rm eff}$ of a "Big Data" set

Equate its MSE to that from a SRS with size $n_{\rm eff}$:

$$D_I \left[\frac{N-n}{n}\right]\sigma^2 = \frac{1}{N-1}\left[\frac{N-n_{\rm eff}}{n_{\rm eff}}\right]\sigma^2$$

What matters is the relative size $f = n/N$

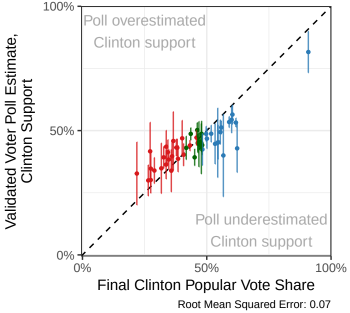
$$n_{\rm eff} = \frac{n}{1+(1-f)[(N-1)D_I-1]} \approx \frac{f}{1-f}\frac{1}{\hat{\rho}^2}.$$

CCES: **Cooperative Congressional Election Study**

(Conducted by Stephen Ansolabehere, Brian Schaffner, Sam Luks, Douglas Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki)

CCES: **Cooperative Congressional Election Study**

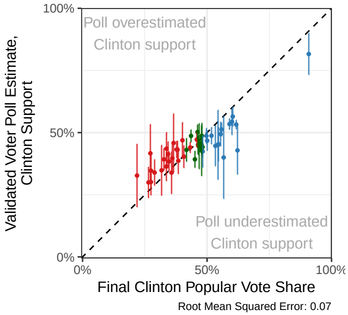
(Conducted by Stephen Ansolabehere, Brian Schaffner, Sam Luks, Douglas Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki)

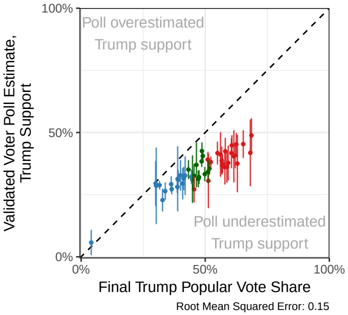


Reasonable predictions for Clinton's Vote Share

CCES: **Cooperative Congressional Election Study**

(Conducted by Stephen Ansolabehere, Brian Schaffner, Sam Luks, Douglas Rivers on **Oct 4 - Nov 6, 2016** (YouGov); Analysis assisted by Shiro Kuriwaki)





Reasonable predictions for Clinton's Vote Share

Serious underestimation of Trump's Vote Share

Assessing $\hat{\rho}$ using state-level data

Xiao-Li Meng
Department of Statistics, Harvard University

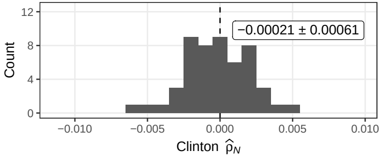
Let μ_N be the true share, and $\hat{\mu}_n$ the estimated share. Then

$$\hat{\rho} = \frac{\hat{\mu}_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N(1-\mu_N)$$

Assessing $\hat{\rho}$ using state-level data

Let μ_N be the true share, and $\hat{\mu}_n$ the estimated share. Then

$$\hat{\rho} = \frac{\hat{\mu}_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N(1-\mu_N)$$



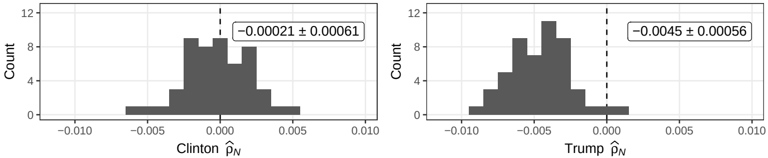
Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$

Assessing $\hat{\rho}$ using state-level data

Xiao-Li Meng
Department of Statistics, Harvard University

Let μ_N be the true share, and $\hat{\mu}_n$ the estimated share. Then

$$\hat{\rho} = \frac{\hat{\mu}_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N(1-\mu_N)$$



Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$ Trump: $\hat{\rho} \approx -0.0045 \pm 0.0006$

What's the implication of $\hat{\rho} = -0.005$?

Xiao-Li Meng
Department of Statistics, Harvard University

- Many (major) survey results published before Nov 8, 2016;

What's the implication of $\hat{\rho} = -0.005$?

Xiao-Li Meng
Department of Statistics, Harvard University

- Many (major) survey results published before Nov 8, 2016;
- Roughly amounts to 1% of eligible voters: $n \approx 2,300,000$;
- Equivalent to 2,300 surveys of 1,000 respondents each.

What's the implication of $\hat{\rho} = -0.005$?

- Many (major) survey results published before Nov 8, 2016;
- Roughly amounts to 1% of eligible voters: $n \approx 2,300,000$;
- Equivalent to 2,300 surveys of 1,000 respondents each.

When $\hat{\rho} = -0.005 = -1/200$, $D_I = 1/40000$, and hence

$$n_{\text{eff}} = \frac{f}{1-f}\frac{1}{D_I} = \frac{1}{99} \times 40000 \approx 404!$$

Xiao-Li Meng
Department of Statistics, Harvard University

What's the implication of $\hat{\rho} = -0.005$?

- Many (major) survey results published before Nov 8, 2016;
- Roughly amounts to 1% of eligible voters: $n \approx 2,300,000$;
- Equivalent to 2,300 surveys of 1,000 respondents each.

When $\hat{\rho} = -0.005 = -1/200$, $D_I = 1/40000$, and hence

$$n_{\text{eff}} = \frac{f}{1-f}\frac{1}{D_I} = \frac{1}{99} \times 40000 \approx 404!$$

- **A** 99.98% **reduction in** n**, caused by** $\hat{\rho} = -0.005$**.**

What's the implication of $\hat{\rho} = -0.005$?

- Many (major) survey results published before Nov 8, 2016;
- Roughly amounts to 1% of eligible voters: $n \approx 2,300,000$;
- Equivalent to 2,300 surveys of 1,000 respondents each.

When $\hat{\rho} = -0.005 = -1/200$, $D_I = 1/40000$, and hence

$$n_{\text{eff}} = \frac{f}{1-f}\frac{1}{D_I} = \frac{1}{99} \times 40000 \approx 404!$$

- **A** 99.98% **reduction in** n, **caused by** $\hat{\rho} = -0.005$.
- **Butterfly Effect** due to Law of Large Populations (LLP)

$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\,\hat{\rho}$$

What's the implication of $\hat{\rho} = -0.005$?

- Many (major) survey results published before Nov 8, 2016;
- Roughly amounts to 1% of eligible voters: $n \approx 2,300,000$;
- Equivalent to 2,300 surveys of 1,000 respondents each.

When $\hat{\rho} = -0.005 = -1/200$, $D_I = 1/40000$, and hence

$$n_{\text{eff}} = \frac{f}{1-f}\frac{1}{D_I} = \frac{1}{99} \times 40000 \approx 404!$$

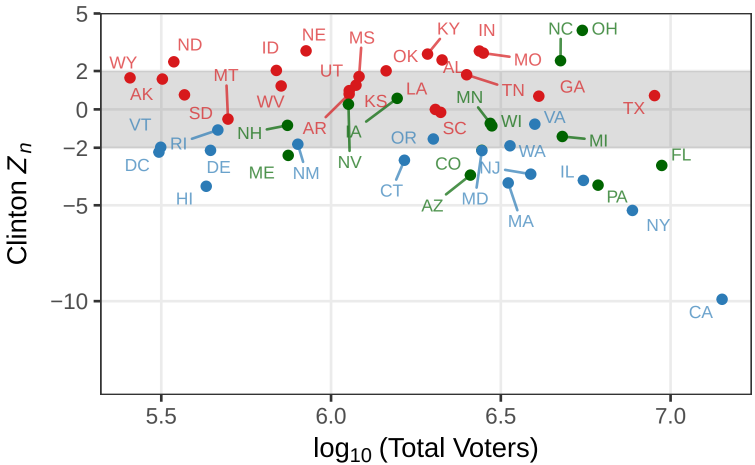
- **A** 99.98% **reduction in** n**, caused by** $\hat{\rho} = -0.005$.
- **Butterfly Effect** due to Law of Large Populations (LLP)

$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\,\hat{\rho}$$

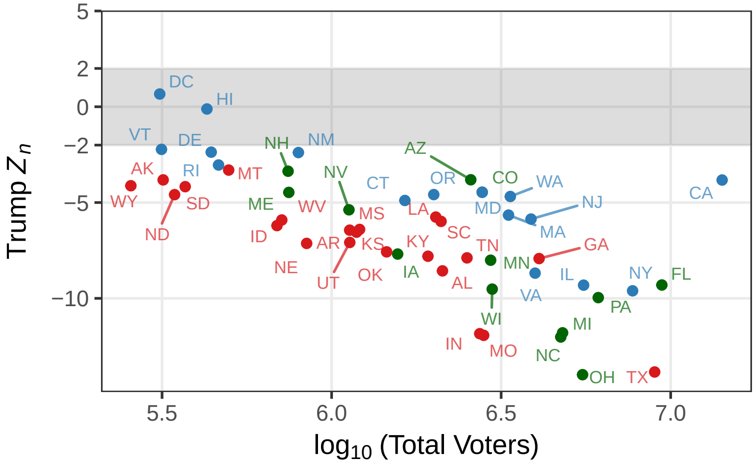
- For $N = 230,000,000$

$$\textbf{Relative Error} = -75.8$$

Visualizing LLP: Actual Coverage for Clinton



Visualizing LLP: Actual Coverage for Trump



Xiao-Li Meng
Department of Statistics, Harvard University

The Big Data Paradox:

If we do not pay attention to data quality, then

The bigger the data,

the surer we fool ourselves.

Lessons Learned ...

- Lesson 1: **What matters most is the quality, not the quantity.**

Lessons Learned ...

- Lesson 1: **What matters most is the quality, not the quantity.**
- Lesson 2: **Don't ignore seemingly tiny probabilistic datasets when combining data sources.**

Lessons Learned ...

Xiao-Li Meng
Department of Statistics, Harvard University

- Lesson 1: **What matters most is the quality, not the quantity.**
- Lesson 2: **Don't ignore seemingly tiny probabilistic datasets when combining data sources.**
- Lesson 3: **Watch the relative size, not the absolute size.**

Lessons Learned ...

Xiao-Li Meng
Department of Statistics, Harvard University

- Lesson 1: **What matters most is the quality, not the quantity.**
- Lesson 2: **Don't ignore seemingly tiny probabilistic datasets when combining data sources.**
- Lesson 3: **Watch the relative size, not the absolute size.**
- Lesson 4: **Probabilistic sampling is an extremely powerful tool to ensure data quality, but it is not the only strategy.**

Lessons Learned ...

Xiao-Li Meng
Department of Statistics, Harvard University

- Lesson 1: **What matters most is the quality, not the quantity.**
- Lesson 2: **Don't ignore seemingly tiny probabilistic datasets when combining data sources.**
- Lesson 3: **Watch the relative size, not the absolute size.**
- Lesson 4: **Probabilistic sampling is an extremely powerful tool to ensure data quality, but it is not the only strategy.**
- Lesson 5: **We may all have had too much "confidence" in big size ...**

Xiao-Li Meng
Department of Statistics, Harvard University

More Lessons From ...

HARVARD DATA SCIENCE REVIEW

A Telescopic, Microscopic, and Kaleidoscopic View of Data Science