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**. *Annals of Applied Statistics*

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**. *Annals of Applied Statistics*
- Many thanks to **Stephen Ansolabehere and Shiro Kuriwaki** for the CCES (**Cooperative Congressional Election Study**) data and analysis on 2016 US election.

Motivating questions

Xiao-Li Meng
Department of Statistics, Harvard University

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivating questions

- 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%?** (Wu, 2012, Seminar at Harvard Statistics)

Motivating questions

- 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%?** (Wu, 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)

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

- Landmark paper: Jerzy Neyman (1934)

A Bit of History: Theory and Practice

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

A Bit of History: Theory and Practice

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

No need to worry about the population size at all?

- Think about tasting soup ...

No need to worry about the population size at all?

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

No need to worry about the population size at all?

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!**

No need to worry about the population size at all?

- 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

No need to worry about the population size at all?

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

$\Longleftrightarrow$

So where were you on 29.02.18?

Menu 5
Xiao-Li Meng Department of Statistics, Harvard University
Motivation
Soup
Euler Identity
Derivation
Trio
LLP
Whats Big?
CCES
Assessing d.d.i
Paradox
Lessons

So where were you on 29.02.18?

So where were you on 29.02.18?

So where were you on 29.02.18?

- 5 most fundamental numbers in mathematics:

$$0, 1, e, \pi, i = \sqrt{-1}$$

- The unexpected one: $i = \sqrt{-1}$

A statistical counterpart of the Euler's identity?

Xiao-Li Meng
Department of Statistics, Harvard University

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \ldots\}$

A statistical counterpart of the Euler's identity?

Xiao-Li Meng
Department of Statistics, Harvard University

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \dots\}$
- σ: **Standard Deviation** $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \ldots\}$
- σ: **Standard Deviation** $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$
- ρ: **Correlation** $\text{Ave}\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - \text{Ave}(\frac{X_j}{\sigma_x})\text{Ave}(\frac{Y_j}{\sigma_y})$

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\mathrm{Ave}\{X_j, j = 1, \ldots\}$
- σ: **Standard Deviation** $\sqrt{\mathrm{Ave}\{(X_j - \mu)^2\}}$
- ρ: **Correlation** $\mathrm{Ave}\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - \mathrm{Ave}(\frac{X_j}{\sigma_x})\mathrm{Ave}(\frac{Y_j}{\sigma_y})$
- n: **Sample Size**

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \ldots\}$
- σ: **Standard Deviation** $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$
- ρ: **Correlation** $\text{Ave}\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - \text{Ave}(\frac{X_j}{\sigma_x})\text{Ave}(\frac{Y_j}{\sigma_y})$
- n: **Sample Size**
- N: **Population Size** The unexpected one ...

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \ldots\}$
- σ: **Standard Deviation** $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$
- ρ: **Correlation** $\text{Ave}\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - \text{Ave}(\frac{X_j}{\sigma_x})\text{Ave}(\frac{Y_j}{\sigma_y})$
- n: **Sample Size**
- N: **Population Size** The unexpected one ...

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

- μ: **Average/Mean** $\text{Ave}\{X_j, j = 1, \ldots\}$
- σ: **Standard Deviation** $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$
- ρ: **Correlation** $\text{Ave}\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - \text{Ave}(\frac{X_j}{\sigma_x})\text{Ave}(\frac{Y_j}{\sigma_y})$
- n: **Sample Size**
- N: **Population Size** The unexpected one ...

The Most Beautiful Statistical Identity?

$$\hat{\mu}_n - \mu_N = \hat{\rho}\sigma\sqrt{\frac{N-n}{n}}$$

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

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

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:

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

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:

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

Actual estimation error

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

Xiao-Li Meng
Department of Statistics, Harvard University

Data quality, quantity, and uncertainty

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

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

Data quality, quantity, and uncertainty

Menu 8

Xiao-Li Meng
Department of Statistics, Harvard University

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

Xiao-Li Meng
Department of Statistics, Harvard University

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): $D_I = \mathsf{E}_R(\hat{\rho}^2)$

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: $D_I \propto N^{-1}$
- Deep trouble when D_I does not vanish with N^{-1} or equivalently $\hat{\rho}$ with $N^{-1/2}$ ...

A Law of Large Populations (LLP)

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

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

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 $\uparrow \sqrt{N}$:

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

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)

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

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

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

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

What's Big? Relative Size or Absolute Size?

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

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

where $f = n/N$ is the **relative size**.

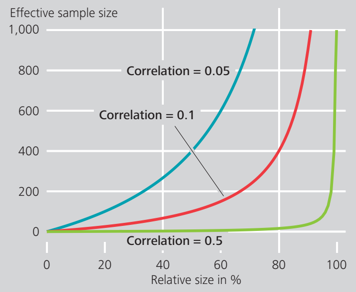
What's Big? Relative Size or Absolute Size?

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

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

where $f = n/N$ is the **relative size**.

The effective sample size of a "Big Data" in terms of SRS size



Deutsche Bundesbank

- If $\hat{\rho} = 0.05$, then $n_{\text{eff}} = 400$ when $f = 1/2$.

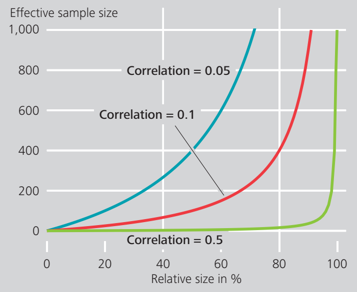
What's Big? Relative Size or Absolute Size?

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

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

where $f = n/N$ is the **relative size**.

The effective sample size of a "Big Data" in terms of SRS size



Deutsche Bundesbank

- If $\hat{\rho} = 0.05$, then $n_{\text{eff}} = 400$ when $f = 1/2$.
- But $f = 1/2$ corresponds to $n \approx 160,000,000$ for the U.S. population;

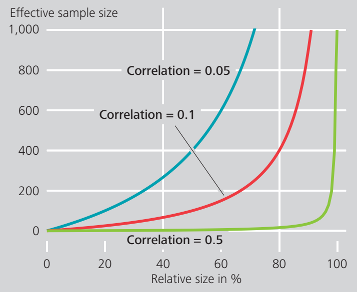
What's Big? Relative Size or Absolute Size?

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

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

where $f = n/N$ is the **relative size**.

The effective sample size of a "Big Data" in terms of SRS size

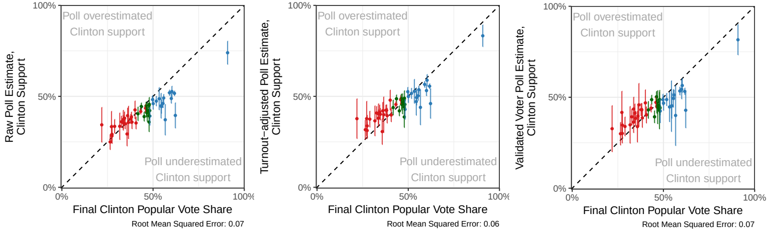


Deutsche Bundesbank

- If $\hat{\rho} = 0.05$, then $n_{\text{eff}} = 400$ when $f = 1/2$.
- But $f = 1/2$ corresponds to $n \approx 160,000,000$ for the U.S. population;
- Hence $\hat{\rho} = 0.05$ implies 99.99975% loss of sample size!

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)

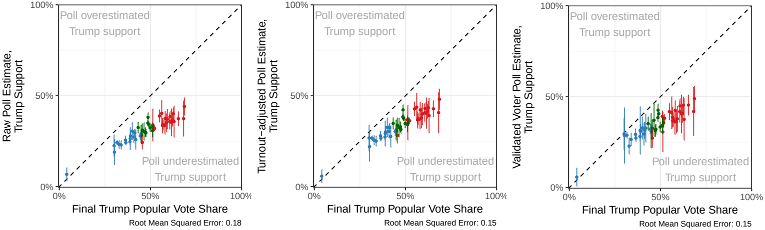


Raw Sample: 64,600   Voting Adj: 48,106   Validated: 34,156

Reasonable predictions for Clinton's Vote Share

Gross under-prediction/reporting of Trump's Share

CCES: **Cooperative Congressional Election Study**



Raw Sample: 64,600 Voting Adj: 48,106 Validated: 34,156

There are many "undecided" ...

Assessing $\hat{\rho}$ using raw counts

Xiao-Li Meng
Department of Statistics, Harvard University

Assessing $\hat{\rho}$ using raw counts

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

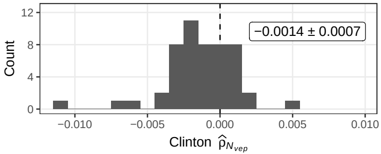
$$\hat{\rho} = \frac{\hat{\mu}_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N(1 - \mu_N)$$

Assessing $\hat{\rho}$ using raw counts

Xiao-Li Meng
Department of Statistics, Harvard University

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

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



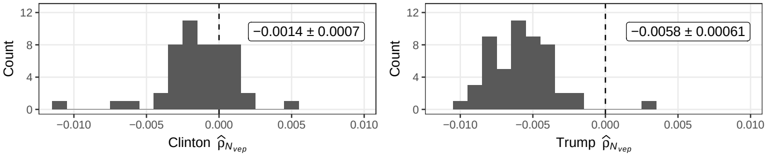
Clinton: $\hat{\rho} \approx -0.0014 \pm 0.0007$

Xiao-Li Meng
Department of Statistics, Harvard University

Assessing $\hat{\rho}$ using raw counts

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

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



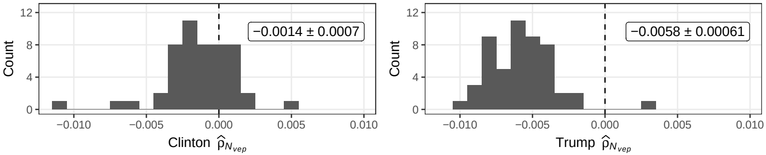
Clinton: $\hat{\rho} \approx -0.0014 \pm 0.0007$ Trump: $\hat{\rho} \approx -0.0058 \pm 0.0006$

Xiao-Li Meng
Department of Statistics, Harvard University

Assessing $\hat{\rho}$ using raw counts

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

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



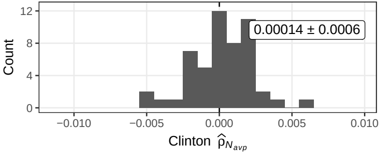
Clinton: $\hat{\rho} \approx -0.0014 \pm 0.0007$ Trump: $\hat{\rho} \approx -0.0058 \pm 0.0006$

- Problem: The mis-match of the *sampled population* and the *actual voting population*

Assessing $\hat{\rho}$ using voting propensity adjusted counts

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

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



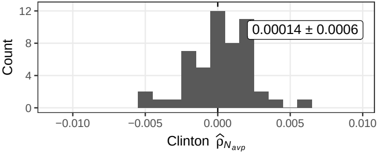
Clinton: $\hat{\rho} \approx 0.0001 \pm 0.0006$

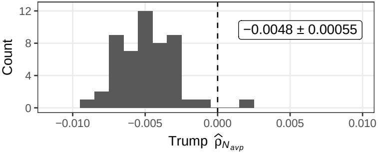
Assessing $\hat{\rho}$ using voting propensity adjusted counts

Xiao-Li Meng
Department of Statistics, Harvard University

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

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



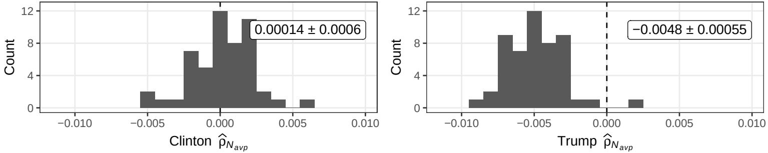


Clinton: $\hat{\rho} \approx 0.0001 \pm 0.0006$ Trump: $\hat{\rho} \approx -0.0048 \pm 0.0005$

Assessing $\hat{\rho}$ using voting propensity adjusted counts

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: $\hat{\rho} \approx 0.0001 \pm 0.0006$ Trump: $\hat{\rho} \approx -0.0048 \pm 0.0005$

- Problem: Estimating voting propensity (and N) is known to be unreliable; weighting may also introduce bias in assessing $\hat{\rho}$.

Assessing $\hat{\rho}_N$ using validated voter counts

Xiao-Li Meng
Department of Statistics, Harvard University

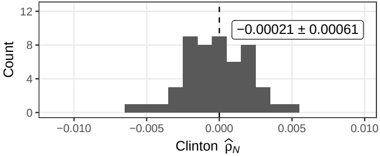
Let μ_N be the true share, and $\hat{\mu}_n$ the estimated share. Then

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

Assessing $\hat{\rho}_N$ using validated voter counts

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

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

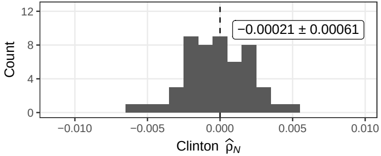


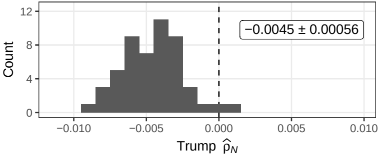
Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$

Assessing $\hat{\rho}_N$ using validated voter counts

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

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



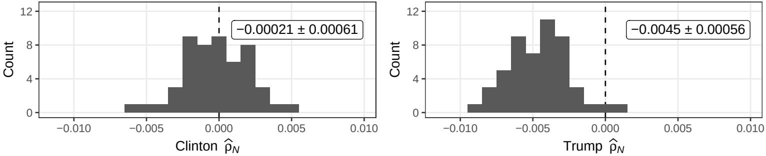


Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$ Trump: $\hat{\rho} \approx -0.0045 \pm 0.0006$

Assessing $\hat{\rho}_N$ using validated voter counts

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

- Problem: Voter validation is done through matching algorithms and it is not fool-proof, and it may introduce additional *selection bias*.

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

Xiao-Li Meng
Department of Statistics, Harvard University

- Many (major) election survey results were published daily for several months before Nov 8, 2016;

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

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) 1% of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 responses each.

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

Xiao-Li Meng
Department of Statistics, Harvard University

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) 1% of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 responses each.

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

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

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

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) 1% of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 responses 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 $\hat{\rho} = -0.005$?

Xiao-Li Meng
Department of Statistics, Harvard University

- Many (major) election survey results were published daily for several months before Nov 8, 2016;
- Roughly amounts to having opinions from (up to) 1% of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 responses each.

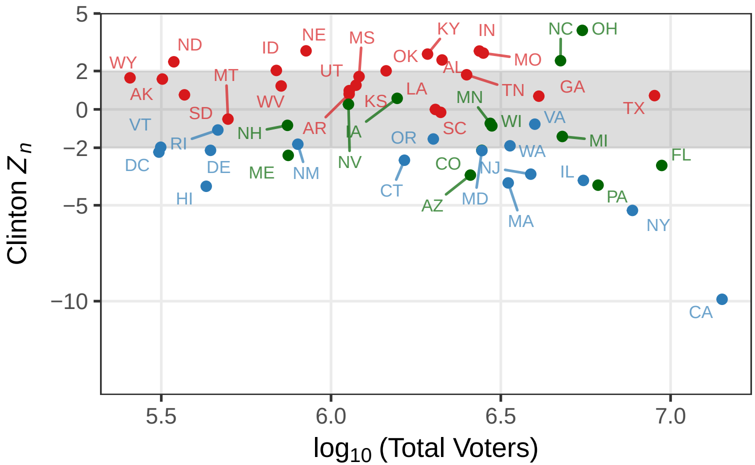
When $\hat{\rho} = -0.005 = -1/200$, $D_I = 1/40000$, and hence

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

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

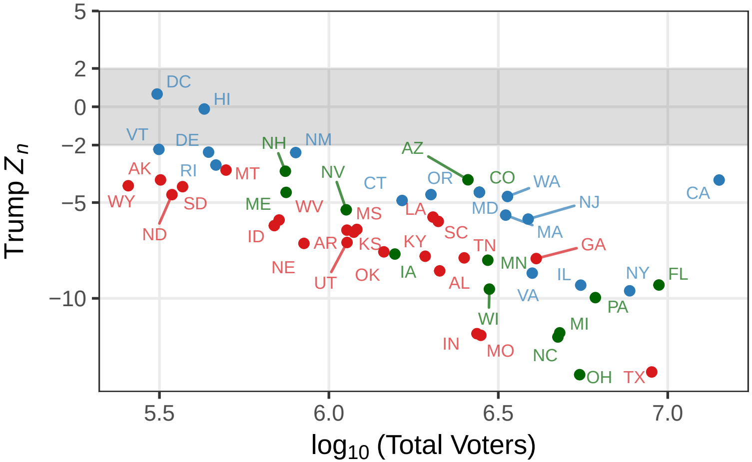
$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\,\hat{\rho}$$

Xiao-Li Meng
Department of Statistics, Harvard University



Visulizing 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.

Xiao-Li Meng
Department of Statistics, Harvard University

Lessons Learned ...

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

Xiao-Li Meng
Department of Statistics, Harvard University

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 ...

- 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 ...

- 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 ...

- 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

… and learning from real experts …

19 things we learned from the 2016 election*

Andrew Gelman† Julia Azari‡

12 July 2017

We can all agree that the presidential election result was a shocker. According to news reports, even the Trump campaign team was stunned to come up a winner.

So now seems like a good time to go over various theories floating around in political science and political reporting and see where they stand, now that this turbulent political year has drawn to a close. In the present article, we go through several things that we as political observers and political scientists have learned from the election, and then discuss implications for the future.

The shock

Immediately following the election there was much talk about the failure of the polls: Hillary Clinton was seen as the clear favorite for several months straight, and then she lost. After all the votes were counted, though, the view is slightly different: by election eve, the national polls were giving Clinton 52 or 53% of the two-party vote, and she ended up receiving 51%. An error of 2 percentage points is no great embarrassment.

The errors in the polls were, however, not uniform. As Figures 1 and 2 show, the Republican candidate outperformed by about 5% in highly Republican states, 2% in swing states, and not at all, on average, in highly Democratic states. This was unexpected in part because, in other recent elections, the errors in poll-based forecasts did not have this sort of structure. In 2016, though, Donald Trump won from his better-than-expected performance in Wisconsin, Michigan, North Carolina, Pennsylvania, and several other swing states.

Trump's win in the general election, and the corresponding success of Republican candidates for the U.S. Senate, then raises two questions: (1) What did the polls get wrong in these key states?, (2) How did Trump and his fellow Republicans do so well? The first is a question about survey respondents, the second a question about voters.

Going backward in time from the election-day shocker, there is the question of how Trump, as a