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 and Xie (2014). **I Got More Data, My Model Is More Refined, But My Estimator Is Getting Worse! Am I Just Dumb?** *Econometric Reviews.* 33: 218-250.
- 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.* 2: 685-726

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 and Xie (2014). **I Got More Data, My Model Is More Refined, But My Estimator Is Getting Worse! Am I Just Dumb?** *Econometric Reviews.* 33: 218-250.
- 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.* 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.

“Big Data” is everywhere ...

Xiao-Li Meng
Department of Statistics, Harvard University

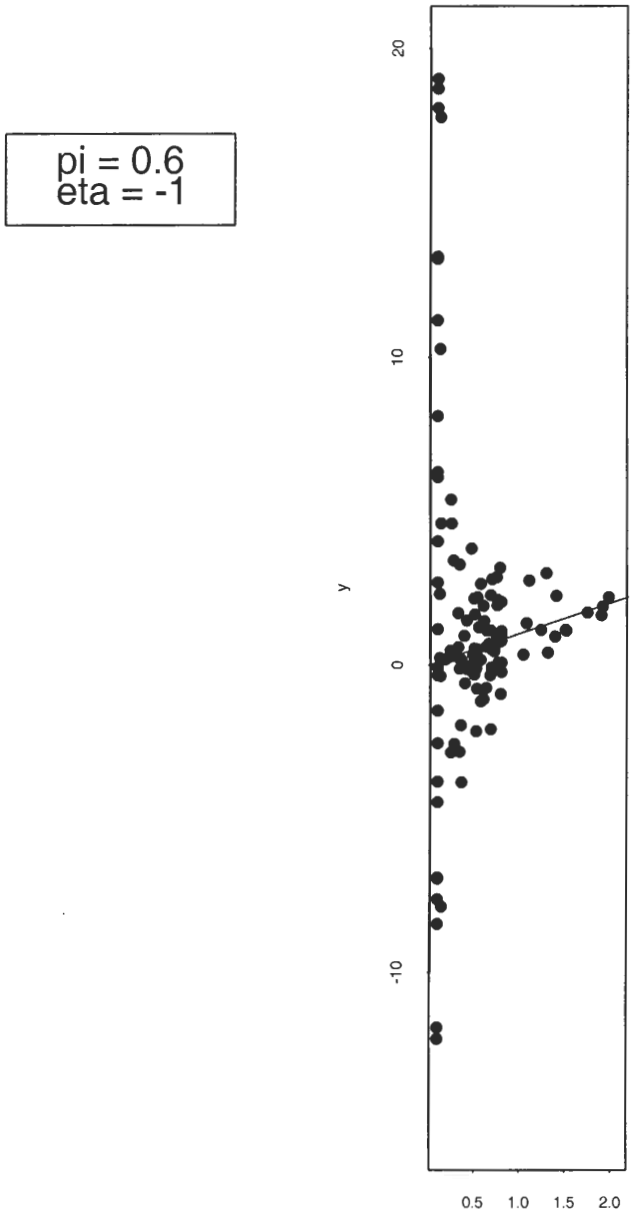
"Big Data" is everywhere ...

Xiao-Li Meng
Department of Statistics, Harvard University

Photo: Sandy Zabell

But Bigger ⇏ Better, even for Least-Squares (LSE)

Xiao-Li Meng
Department of Statistics, Harvard University



Xiao-Li Meng
Department of Statistics, Harvard University

Bigger ⇏ Better: A Mathematical Proof

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^{\eta}), \quad i = 1, \ldots, n$$

Xiao-Li Meng
Department of Statistics, Harvard University

Bigger $\not\Rightarrow$ Better: A Mathematical Proof

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

Xiao-Li Meng
Department of Statistics, Harvard University

Bigger ≠ Better: A Mathematical Proof

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Xiao-Li Meng
Department of Statistics, Harvard University

Bigger ⇏ Better: A Mathematical Proof

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

But $\hat{\beta}^{LSE}$ is not *self-efficient* (Meng, 1994) when $\eta \neq 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{\sum_{i=1}^n X_i^{2+\eta}}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Xiao-Li Meng
Department of Statistics, Harvard University

Bigger ≠ Better: A Mathematical Proof

A Heteroscedastic Regression Model

$$Y_i = \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 X_i^\eta), \quad i = 1, \ldots, n$$

Least-squares estimator:

$$\hat{\beta}^{LSE} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

"Sandwich" estimator of var:

$$\hat{V}^{LSE} = \frac{\sum_{i=1}^n X_i^2 (Y_i - X_i \hat{\beta}^{LSE})^2}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

But $\hat{\beta}^{LSE}$ is not *self-efficient* (Meng, 1994) when $\eta \neq 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{\sum_{i=1}^n X_i^{2+\eta}}{\left[\sum_{i=1}^n X_i^2\right]^2}$$

Compare, when $\eta = 0$:

$$\mathrm{V}(\hat{\beta}^{LSE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^2}$$

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Re-weight the data by $W_i = X_i^{-\eta/2}$ (assume η is known)

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Re-weight the data by $W_i = X_i^{-\eta/2}$ (assume η is known)

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Re-weight the data by $W_i = X_i^{-\eta/2}$ (assume η is known)

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Re-weight the data by $W_i = X_i^{-\eta/2}$ (assume η is known)

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

So it is justifiable to throw away some data points if you don't know how to use them most effectively because

Xiao-Li Meng
Department of Statistics, Harvard University

What went wrong?

- Those observations with large variabilities received more weight than they deserve.

Re-weight the data by $W_i = X_i^{-\eta/2}$ (assume η is known)

$$W_i Y_i = \beta(W_i X_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$\hat{\beta}_{MLE} = \frac{\sum_{i=1}^n X_i^{1-\eta} Y_i}{\sum_{i=1}^n X_i^{2-\eta}}$$

$$\mathrm{V}(\hat{\beta}_{MLE} | X, \theta) = \sigma^2 \frac{1}{\sum_{i=1}^n X_i^{2-\eta}}$$

So it is justifiable to throw away some data points if you don't know how to use them most effectively because

When the optimal W_i's have large variation, **setting small W_i's to zero better approximates the optimal weighting scheme than "blindly" using equal weights.**

Don't sue me ...

WHEN IT SEEMS DESIRABLE TO IGNORE DATA

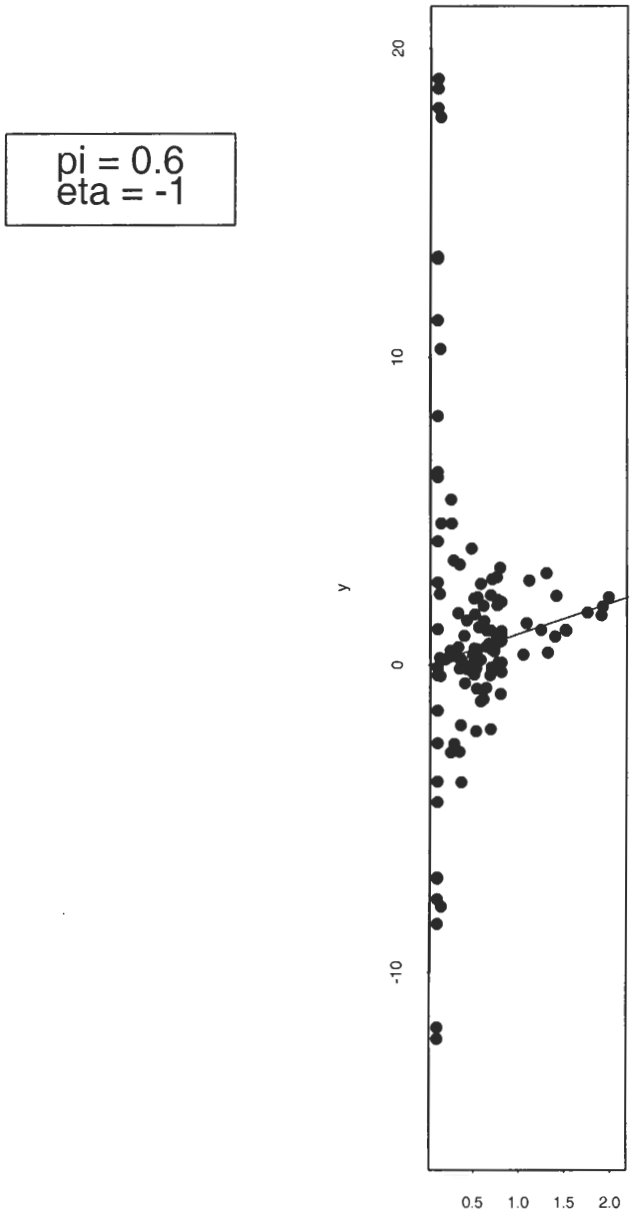
Herman Chernoff
Massachusetts Institute of Technology

ABSTRACT

An experiment designed to detect the relative motion of two astronomical objects raised the problem of testing, against shift alternatives, the hypotehsis H_0 that two energy distributions are equivalent. The relevant data consist of independent Poisson counts X_{ij} with means $\lambda_j p_{ij} T_{ij}$ where λ_j is the intensity of radiation from the jth object, p_{ij} is the probability that a random photon from the jth object has energy in a small interval centered about e_i, and T_{ij} is the time duration allocated to the count X_{ij}. The hypothesis H_o implies that $p_{i1} = p_{i2}$ for i = 1, 2, ..., m.

A natural test uses the statistic $\Sigma e_i(\hat{p}_{i2} - \hat{p}_{i1})$ where the $\hat{p}_{ij}$ are estimates of p_{ij}. For intervals where the p_{ij} were anticipated to be small, the experimenter chose small T_{ij} values

From Robins and Wang (2000, *Biometrika*)



Xiao-Li Meng
Department of Statistics, Harvard University

Data Quality vs Quantity: 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).

Xiao-Li Meng
Department of Statistics, Harvard University

Data Quality vs Quantity: 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)

Xiao-Li Meng
Department of Statistics, Harvard University

Data Quality vs Quantity: 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)

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Bit of History: Theory and Practice

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

- Landmark paper: Jerzy Neyman (1934)

A Bit of History: Theory and Practice

Xiao-Li Meng
Department of Statistics, Harvard University

- **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)

A Bit of History: Theory and Practice

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

No need to worry about the population size at all?

Xiao-Li Meng
Department of Statistics, Harvard University

- Think about tasting soup ...

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?

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?

Xiao-Li Meng
Department of Statistics, Harvard University

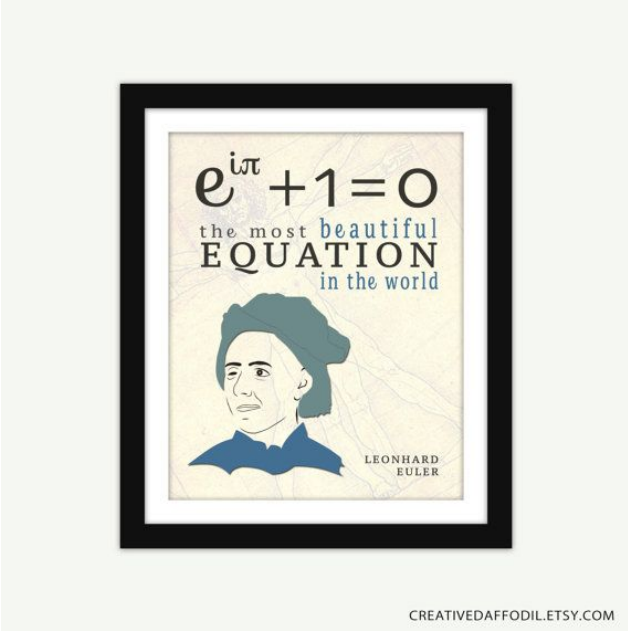
- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

No need to worry about the population size at all?

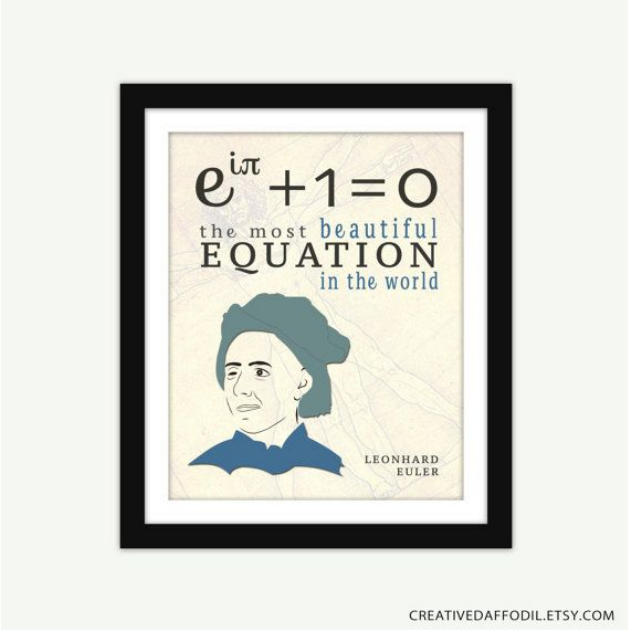
- Think about tasting soup ...
- Stir it well, then a few bits are sufficient **regardless of the size of the container!**

$\Longleftrightarrow$

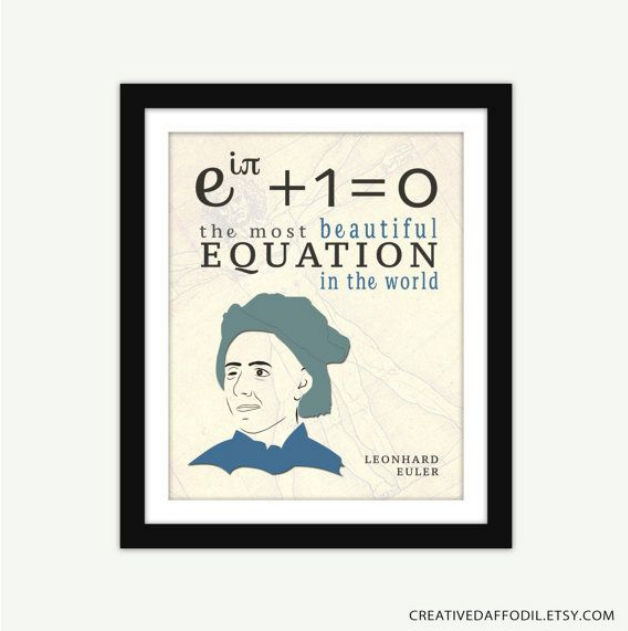
Euler's identity: An unexpected helper



Euler's identity: An unexpected helper



Euler's identity: An unexpected helper

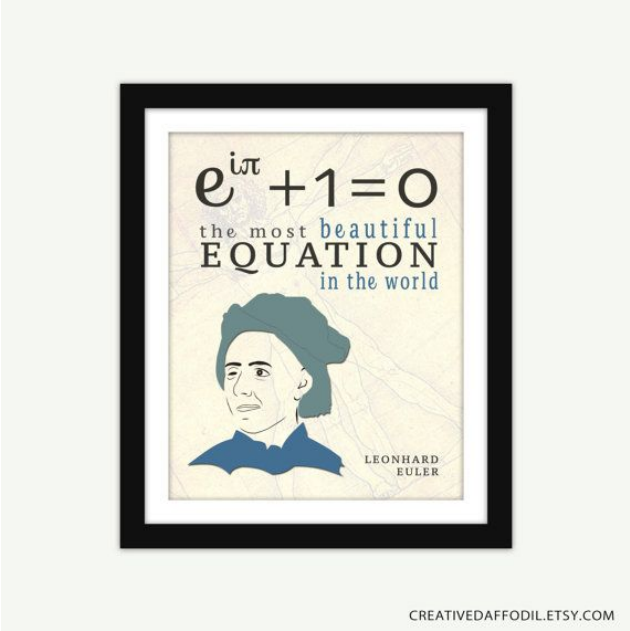


- 5 most fundamental numbers in mathematics:

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

Xiao-Li Meng
Department of Statistics, Harvard University

Euler's identity: An unexpected helper



- 5 most fundamental numbers in mathematics:

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

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

A statistical counterpart of the Euler's identity?

What are the five most fundamental symbols in Statistics?

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

Xiao-Li Meng
Department of Statistics, Harvard University

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?

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

Xiao-Li Meng
Department of Statistics, Harvard University

2016 US Presidential Election

- n: number of respondents to an election survey

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

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

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

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

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:

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

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:

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

Actual estimation error

$$\hat{\mu}_n - \mu_N = \frac{\text{Ave}(R_jX_j)}{\text{Ave}(R_j)} - \text{Ave}(X_j)$$

$$= \left[\frac{\text{Ave}(R_jX_j) - \text{Ave}(R_j)\text{Ave}(X_j)}{\sigma_R\sigma_X}\right] \times \frac{\sigma_R}{\text{Ave}(R_j)} \times \sigma_X$$

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

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

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

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

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

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

Xiao-Li Meng
Department of Statistics, Harvard University

Data Defect Index (d.d.i.)

Mean Squared Error (MSE)

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

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

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

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

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): $D_I = (N-1)^{-1}$
- 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 $N^{-1/2}$ ...

Xiao-Li Meng
Department of Statistics, Harvard University

A Law of Large Populations (LLP)

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

$$\text{“}Z\text{−score”} \equiv \frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

Xiao-Li Meng
Department of Statistics, Harvard University

A Law of Large Populations (LLP)

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

$$\text{“}Z\text{−score”} \equiv \frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

The (lack-of) design effect (Deff)

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

Xiao-Li Meng
Department of Statistics, Harvard University

A Law of Large Populations (LLP)

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

$$\text{“}Z\text{−score”} \equiv \frac{\text{Actual Error}}{\text{Benchmark SRS Standard Error}} = \sqrt{N-1}\hat{\rho}$$

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}:

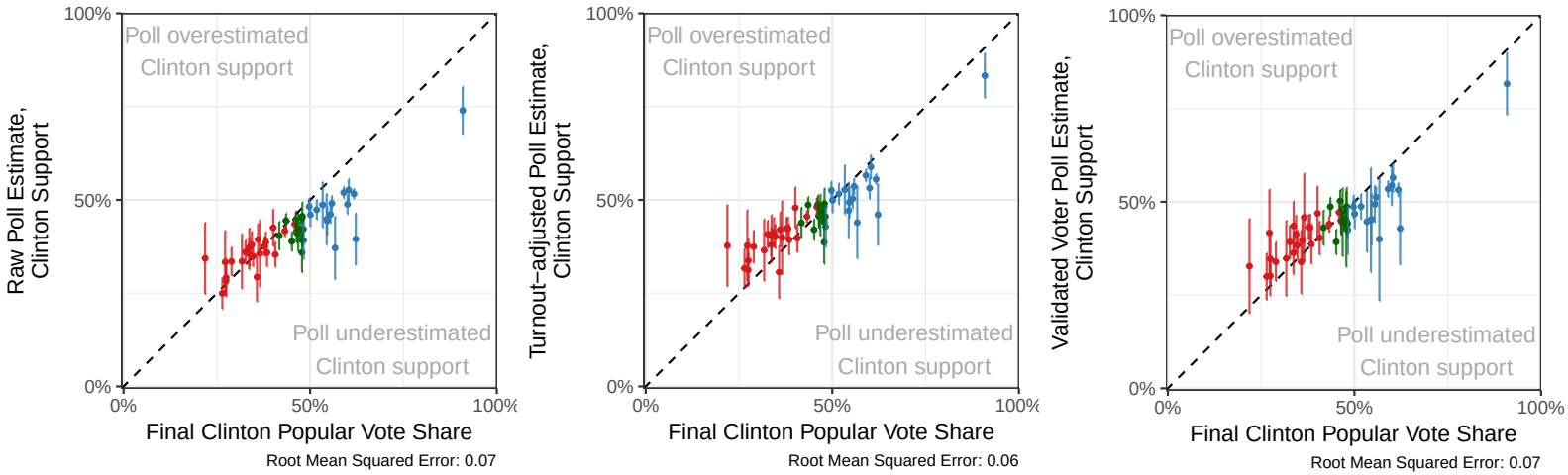
$$D_I \left[\frac{1-f}{f}\right]\sigma^2 = \frac{1}{N-1}\left[\frac{N-n_{\text{eff}}}{n_{\text{eff}}}\right]\sigma^2$$

Xiao-Li Meng
Department of Statistics, Harvard University

Gaining 2020 Vision: Assessing the behavioral $\hat{\rho}$

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)



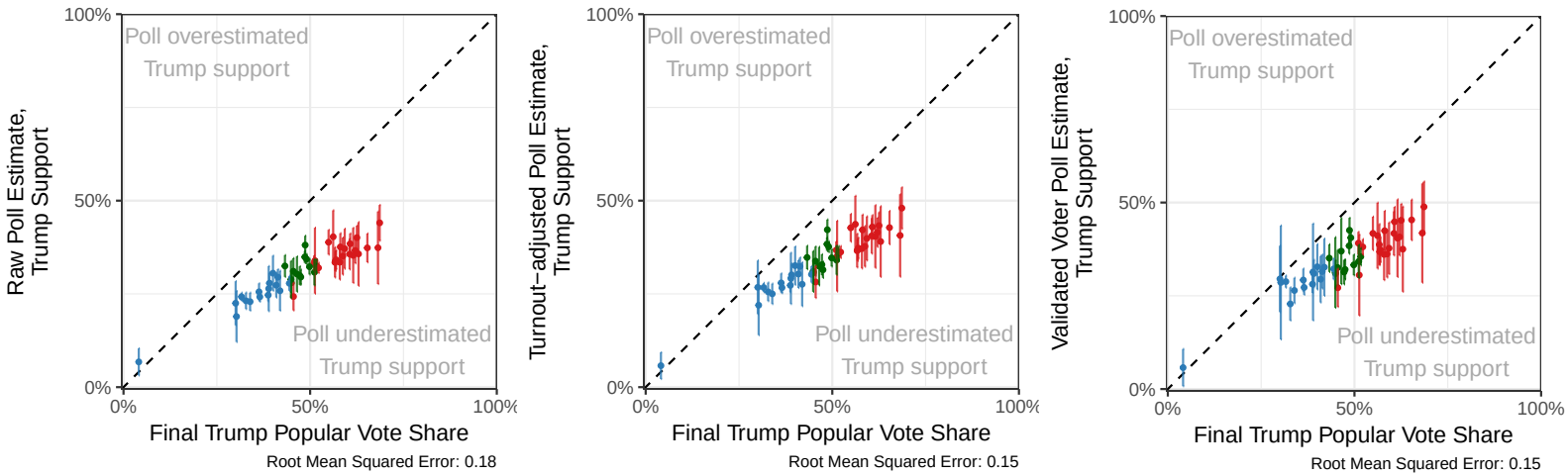
Raw Sample: 64,600 Voting Adj: 48,106 Validated: 34,156

Reasonable predictions for Clinton's Vote Share

Xiao-Li Meng
Department of Statistics, Harvard University

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" ...

Xiao-Li Meng
Department of Statistics, Harvard University

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

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

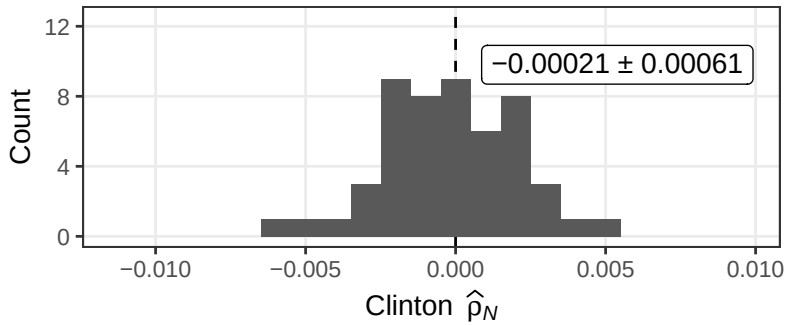
$$\hat{\rho} = \frac{\hat{\mu}_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N(1-\mu_N)$$

Assessing $\hat{\rho}$ 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

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



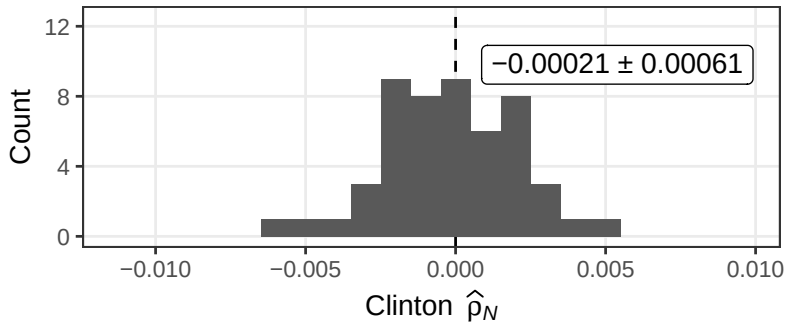
Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$

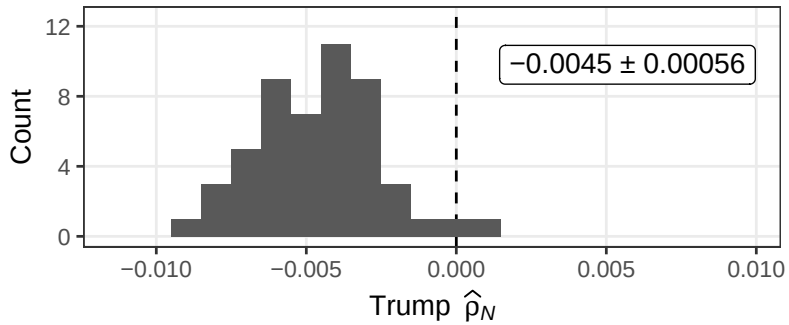
Xiao-Li Meng
Department of Statistics, Harvard University

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

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

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



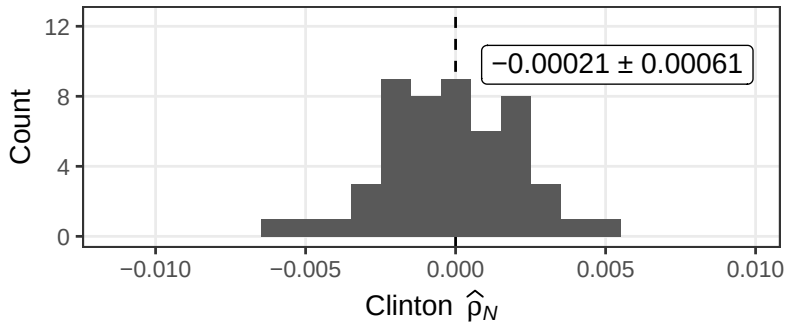


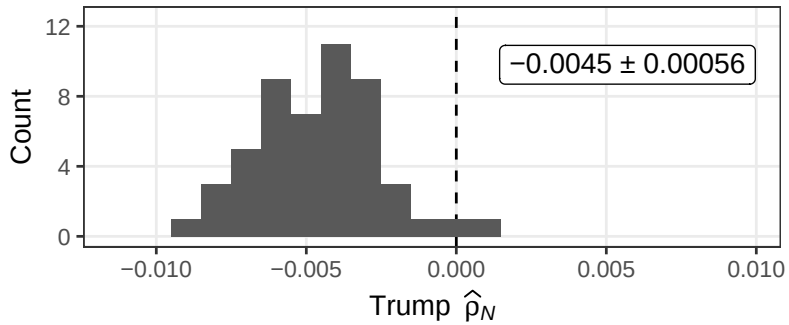
Clinton: $\hat{\rho} \approx -0.0002 \pm 0.0006$ Trump: $\hat{\rho} \approx -0.0045 \pm 0.0006$

Assessing $\hat{\rho}$ 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*.

Xiao-Li Meng
Department of Statistics, Harvard University

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

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

Xiao-Li Meng
Department of Statistics, Harvard University

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) $f = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

Xiao-Li Meng
Department of Statistics, Harvard University

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

Xiao-Li Meng
Department of Statistics, Harvard University

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) $f = 1\%$ of US voting eligible population: $n \approx 2,300,000$;
- Equivalent to about 2,300 surveys of 1,000 respondents each.

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

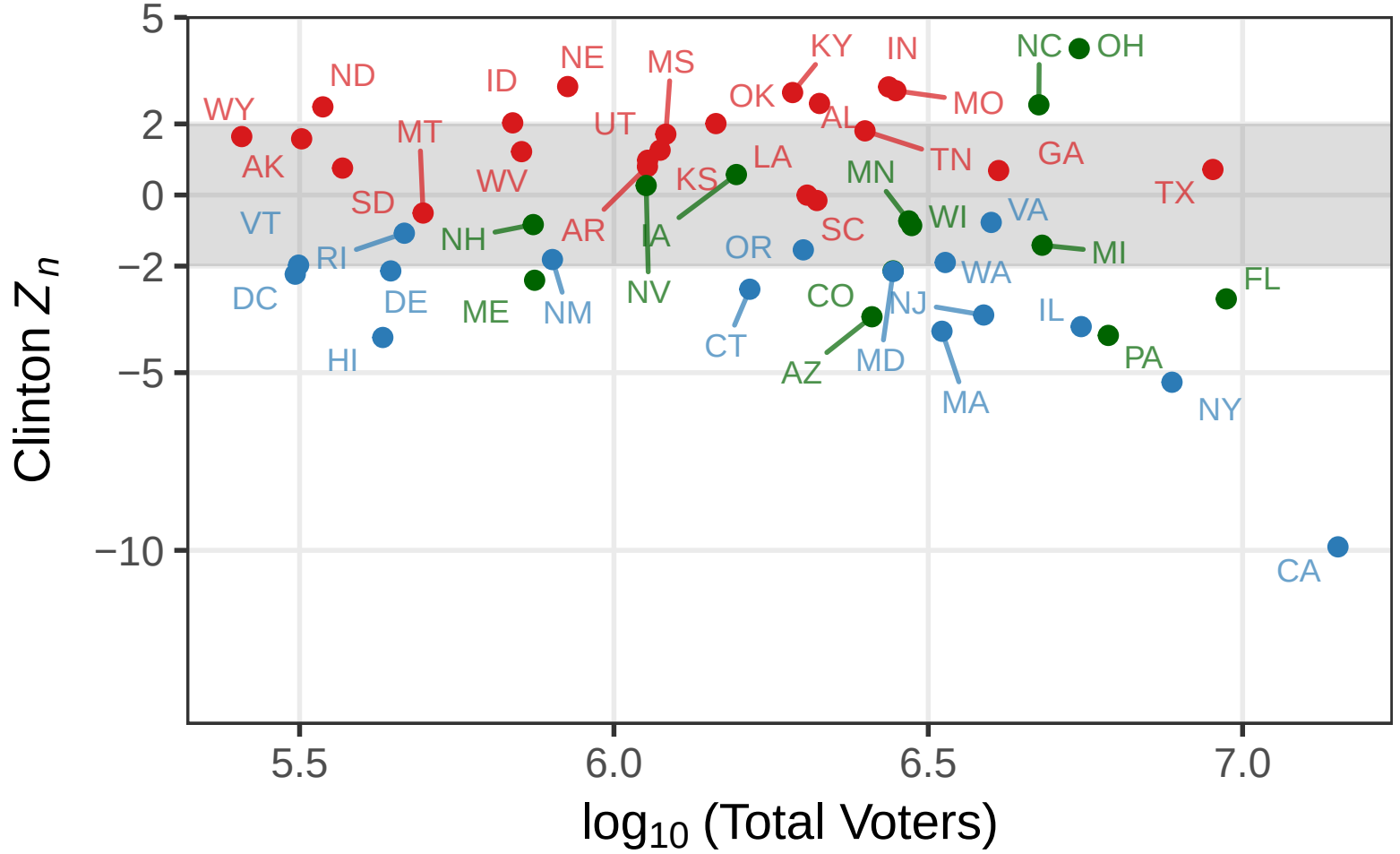
$$n_{\text{eff}} = \frac{f}{1-f}\frac{1}{D_I} = \frac{1}{99} \times 40000 \approx 404!$$

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

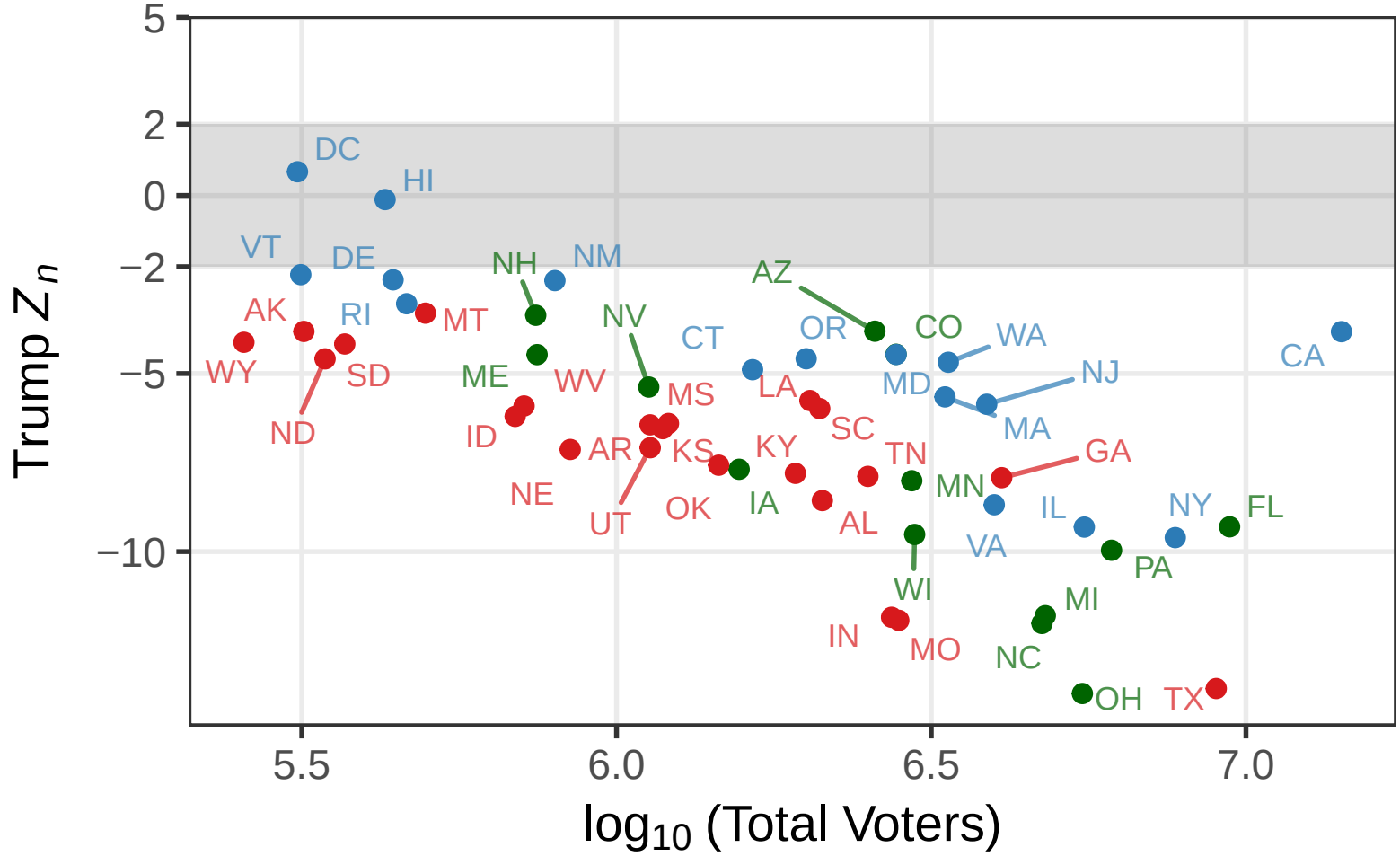
$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\hat{\rho}$$

Visulizing LLP: Actual Coverage for Clinton

Xiao-Li Meng
Department of Statistics, Harvard University



Visulizing LLP: Actual Coverage for Trump



The Big Data Paradox:

Xiao-Li Meng
Department of Statistics, Harvard University

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.**

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.**
- Lesson 3: **Watch the relative size, not the absolute size.**

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.**
- 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.**

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.**
- 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 ...**

... 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