Menu 1

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

"Trio" Identity

Trio

LLP

What's Big?

CCES

Assessing d.d.i

Paradox

Lessons

Statistical Paradises and Paradoxes in Big Data (I):

Law of Large Populations, Big Data Paradox, and the 2016 US Election

Xiao-Li Meng
Department of Statistics, Harvard University

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

Statistical Paradises and Paradoxes in Big Data (I):

Law of Large Populations, Big Data Paradox, and the 2016 US Election

Xiao-Li Meng
Department of Statistics, Harvard University

- Meng (2018). *Annals of Applied Statistics*, No. 2, 685-726.
  `https://statistics.fas.harvard.edu/people/xiao-li-meng`

Menu 1

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

"Trio" Identity

Trio

LLP

What's Big?

CCES

Assessing d.d.i

Paradox

Lessons

Statistical Paradises and Paradoxes in Big Data (I):

Law of Large Populations, Big Data Paradox, and the 2016 US Election

Xiao-Li Meng
Department of Statistics, Harvard University

- Meng (2018). *Annals of Applied Statistics*, No. 2, 685-726. `https://statistics.fas.harvard.edu/people/xiao-li-meng`
- 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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

- 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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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%?** (Jeremy Wu of US Census Bureau, 2012, Seminar at Harvard Statistics)

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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%?** (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)

A Bit of History: Theory and Practice

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

A Bit of History: Theory and Practice

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

- **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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

- **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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

- **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

When/why can we ignore the population size?

Menu 4

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

- Think about tasting soup ...

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

When/why can we ignore the population size?

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

When/why can we ignore the population size?

Menu 4

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

When/why can we ignore the population size?

Menu 4

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

When/why can we ignore the population size?

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

$\Longleftrightarrow$

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

2016 US Presidential Election

- n: number of respondents to an election survey

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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:

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

Data quality, quantity, and uncertainty

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

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

Data quality, quantity, and uncertainty

Menu 6

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

A Law of Large Populations (LLP)

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

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

A Law of Large Populations (LLP)

If $\rho = \mathsf{E}_R(\hat{\rho}) \neq 0$, 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)

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

A Law of Large Populations (LLP)

If $\rho = \mathsf{E}_R(\hat{\rho}) \neq 0$, 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":

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

Effective Sample Size

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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}:

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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}:

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

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

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

Gaining 2020 Vision: Assessing the behavioral $\hat{\rho}$ using validated voter counts ($\approx 35,000$)

Menu 10

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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)

Gaining 2020 Vision: Assessing the behavioral $\hat{\rho}$ using validated voter counts ($\approx 35,000$)

Menu 10

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

"Trio" Identity

Trio

LLP

What's Big?

CCES

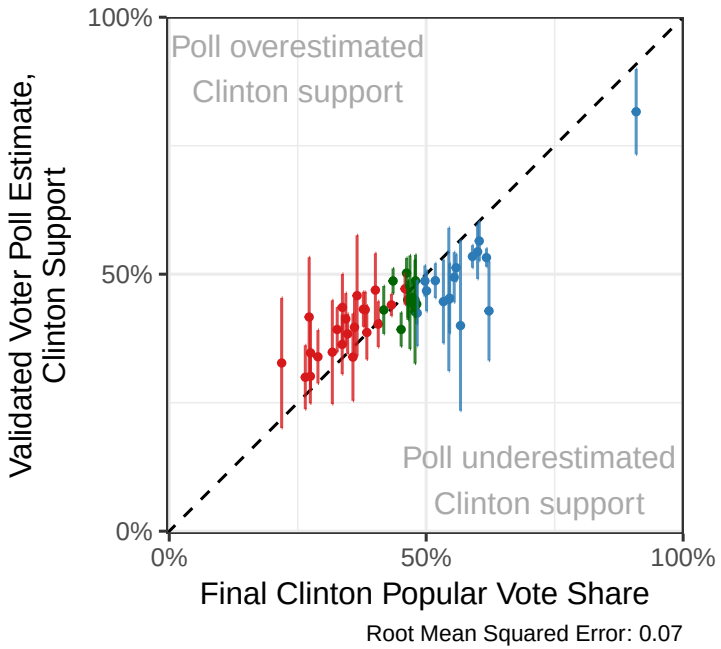
Assessing d.d.i

Paradox

Lessons

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

Gaining 2020 Vision: Assessing the behavioral $\hat{\rho}$ using validated voter counts ($\approx 35,000$)

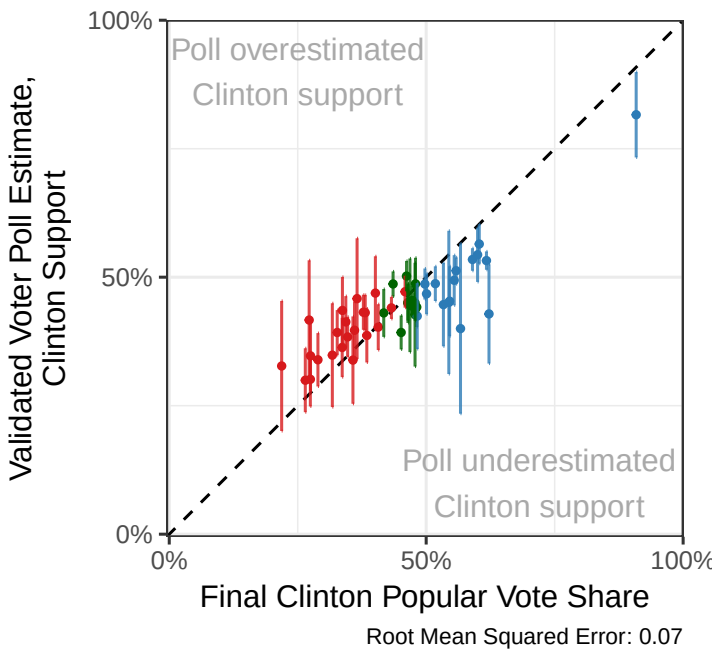
Menu 10

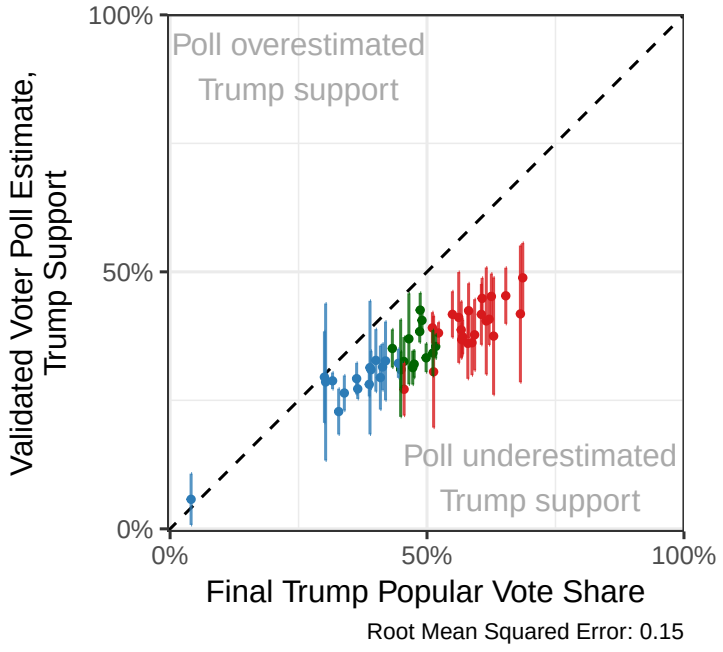
Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

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

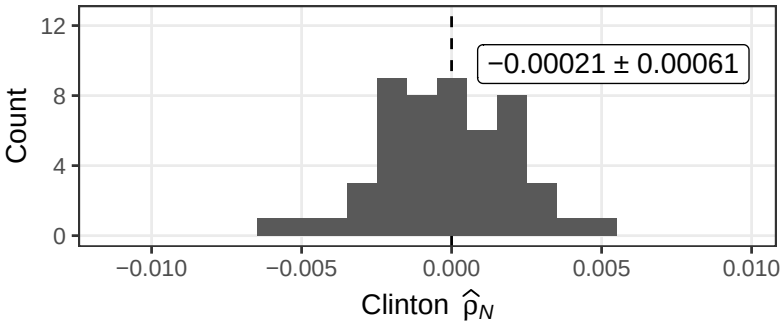
Assessing $\hat{\rho}$ using state-level data

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

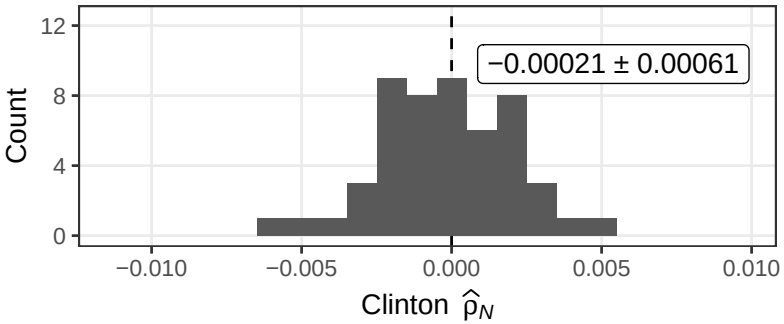
Xiao-Li Meng
Department of Statistics, Harvard University

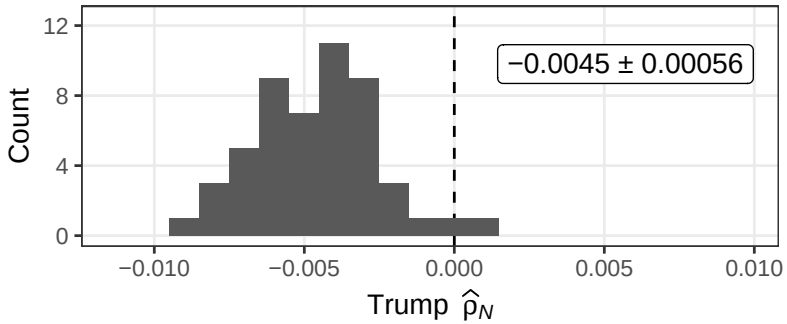
Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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.

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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 $\hat{\rho} = -0.005$.**

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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)

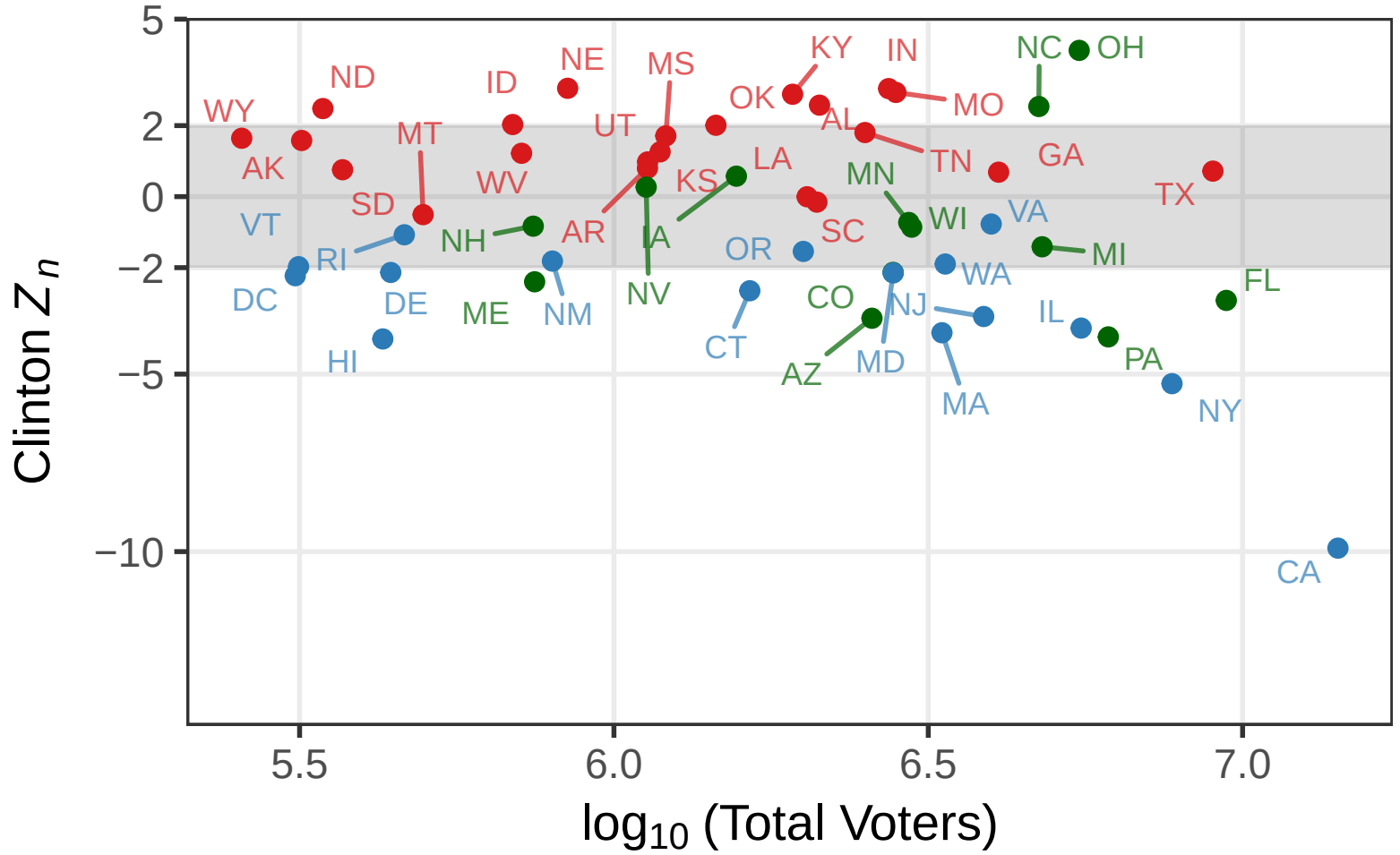
$$\textbf{Relative Error} = \sqrt{\mathbf{N-1}}\hat{\rho}$$

Visualizing LLP: Actual Coverage for Clinton

Menu 13

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

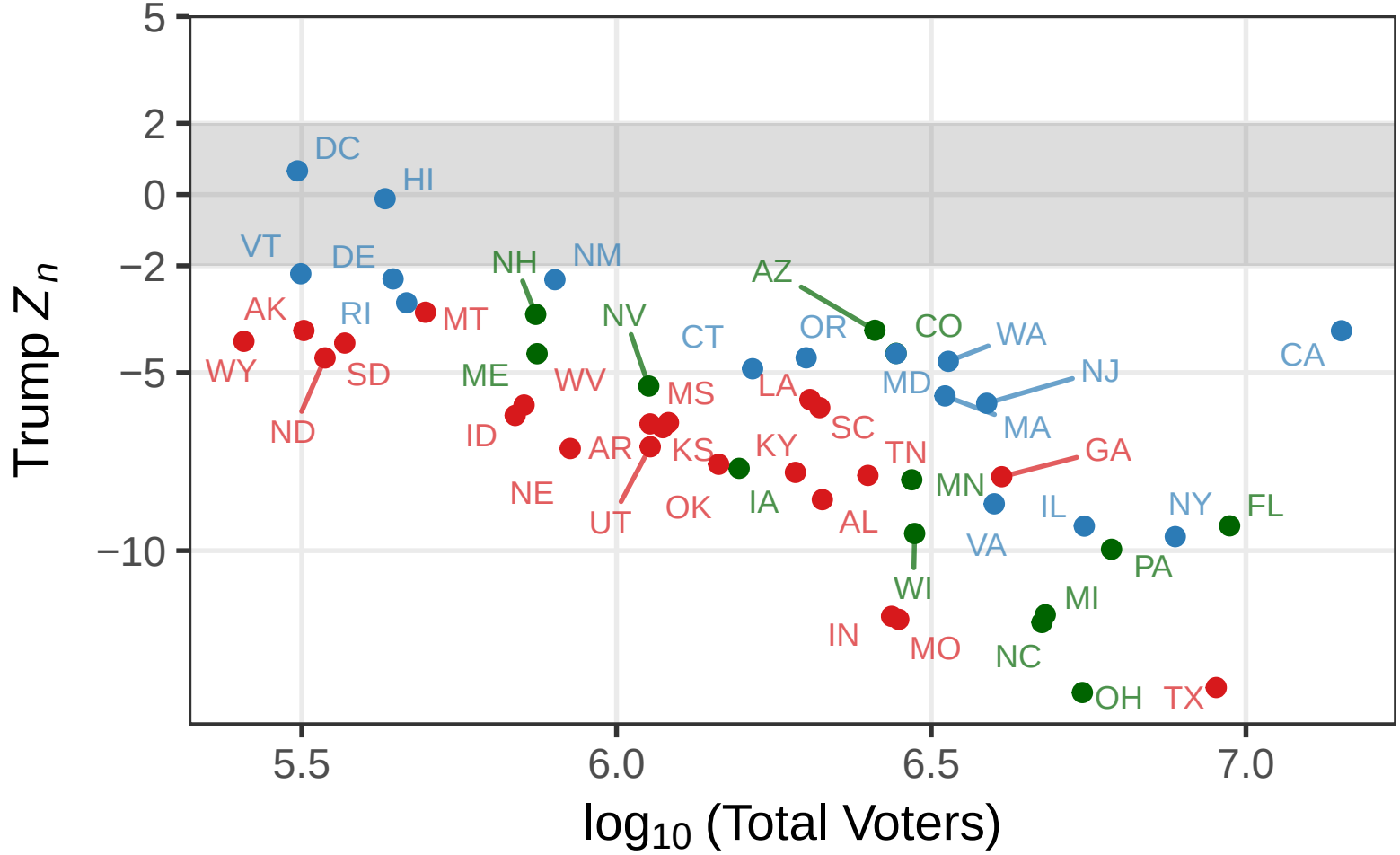


Visualizing LLP: Actual Coverage for Trump

Menu 14

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons



Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

Lessons Learned ...

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

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

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: **Classical theory is BIG for "big data", as long as we let it go outside the classical box.**

In case you are kind enough to invite me again ...

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

Xiao-Li Meng
Department of Statistics, Harvard University

Motivation
Soup
"Trio" Identity
Trio
LLP
What's Big?
CCES
Assessing d.d.i
Paradox
Lessons

In case you are kind enough to invite me again ...

The sequel: Meng (2018/9)

Statistical Paradises and Paradoxes in Big Data (II): Multi-resolution Inference, Simpson's Paradox, and Individualized Treatments