

Menu

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

Xiao-Li Meng Department of Statistics, Harvard University



Menu

Department of 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.

The Annals of Applied Statistics Vol 2: 685-726



Menu

Department of 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.
 - The Annals of Applied Statistics Vol 2: 685-726
- Many thanks to Stephen Ansolabehere and Shiro Kuriwaki for the CCES (Cooperative Congressional **Election Study**) data and analysis on 2016 US election.



A Painful Reminder ... (from internet)

Menu

Department of



The day before the 2016 US Presidential Election. most pollsters and statistical models had pegged Hillary Clinton's chances of winning at greater than 90%.

99%

98%

92%

91%

89%

85%

72%

Princeton Flection Consortium

Huffington Post

Daily KOS

CNN

PredictWise

New York Times



Five Thirtu Eight



Menu

Soup



Menu

•

Xiao-Li Mer Department Statistics, Harvard University

Soup

"Trio" Ident

Trio

ПP

What's Bi

CCES

Assessing d.o

Parado

essons

Think about tasting soup ...



Menu

citu

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identit

LLP

What's Bi

CCES

/ (33C33IIIg

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





Menu

Xiao-Li Meng Department of Statistics, Harvard

Soup

"Trio" Identii

LLP

What's Ri

CCEC

Assessing d.o

Paradox

eccone

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





Menu

Department of

Soup

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







Menu

rienu

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identi

Trio

LLE

What's Big

Assessing d d i

Assessing u.u.i

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









Menu

nu

Xiao-Li Men
Department
Statistics,
Harvard
University

Soup

"Trio" Identity

Tric

LLP

What's Bi

CES

Assessing d.d

Parado

occone

• n: number of respondents to an election survey



Menu

Xiao-Li Men Department of Statistics, Harvard University

Soup

"Trio" Identity

Tric

IIР

What's Bi

CES

Assessing d.d

Parado:

occon

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



Menu

"Trio" Identity

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



Menu

Xiao-Li Men Department Statistics, Harvard University

Sout

"Trio" Identity

IIР

M/L - U - D*

vviiat 3 Dig.

Assessing d.d

Darada

- 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



Menu

Xiao-Li Men Department Statistics, Harvard University

Sout

"Trio" Identity

IIР

M/L - U - D*

vviiat 3 Dig.

Assessing d.d

Darada

- 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



Menu

Xiao-Li Meng Department of Statistics, Harvard

Soup

"Trio" Identity

_.

LLP

What's Big?

Assessing d.o

.

- 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

Estimating Trump's share: $\mu_N = Ave(X_j)$ by sample average:

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



Menu

-Li Meng irtment o

Statistics, Harvard University

Soup

"Trio" Identity

T. . .

LLP

What's Big?

Assessing d.d.i

r ar auox

• 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

Estimating Trump's share: $\mu_N = Ave(X_j)$ by sample average:

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

Actual estimation error

$$\hat{\mu}_{n} - \mu_{N} = \frac{\operatorname{Ave}(R_{j}X_{j})}{\operatorname{Ave}(R_{j})} - \operatorname{Ave}(X_{j})$$

$$= \left[\frac{\operatorname{Ave}(R_{j}X_{j}) - \operatorname{Ave}(R_{j})\operatorname{Ave}(X_{j})}{\sigma_{R}\sigma_{X}}\right] \times \frac{\sigma_{R}}{\operatorname{Ave}(R_{j})} \times \sigma_{X}$$



Data quality, quantity, and uncertainty

Menu

5

Xiao-Li Meng Department o Statistics, Harvard University

Soup

"Trio" Ident

Trio

....

\\/L_+'_ D:

.

Assessing d.d

Parado>

eccone

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

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



Data quality, quantity, and uncertainty

Menu

u

5

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identity

Trio

LLP

What's Ri

vviiat s Dig:

Assessing d.d

Parado

eccons

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

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



Data quality, quantity, and uncertainty

Menu

5

Xiao-Li Meng Department o Statistics, Harvard University

Soup

" Irio" Identity

Trio

LLP

M/L - 12 - D

......

Assessing d.d

Paradox

occon

Because
$$\sigma_R^2 = f(1-f)$$
, $f = \text{Ave}\{R_j\} = \frac{n}{N}$, we have
$$\text{Error} = \underbrace{\hat{\rho}_{R,X}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{Data Quantity}} \times \underbrace{\sigma_X}_{\text{Problem Difficulty}}$$



Menu

1

6

Xiao-Li Meng Department o Statistics, Harvard

Soup

"Trio" Identi

Trio

LLP

What's Big

CCES

Assessing d.d

Parado:

occone

Mean Squared Error (MSE)

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



Menu

6

Xiao-Li Mer Department Statistics, Harvard University

Soup

" Irio" Identi

Trio

ПР

\\/|----

CCES

Assessing d

Parado:

esson

Mean Squared Error (MSE)

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

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

Menu

u

6

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identit

THIS TOURING

Trio

What's Big

CCES

Assessing d.

Parado:

_essons

Mean Squared Error (MSE)

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

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

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

Menu

1

6

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identit

THO Identity

Trio

14/1 J. D.

Wilat's Dig:

Assessing d.

Parado

esson

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): $D_I = E_R(\hat{\rho}^2)$

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

Menu

6

Department of Statistics.

Trio

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): $D_I = E_R(\hat{\rho}^2)$

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



A Law of Large Populations (LLP)

Menu

ieilu

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identii

Trio

LLP

What's Big

CES

Assessing d.d

Parado>

l occon

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

Benchmark SRS Standard Error

$$=\sqrt{N-1}\hat{
ho}$$



A Law of Large Populations (LLP)

Menu

Department of

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)

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



A Law of Large Populations (LLP)

Menu

viellu

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identity

_.

LLP

What's Big

CCES

Assessing d.d.

Loccone

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)

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

Paradigm shift for "Big Data":

$$\frac{\sigma}{\sqrt{n}}$$
random error

to

$$\hat{\rho}\sqrt{N}$$

relative systemtic bias



Effective Sample Size

Menu

8

o-Li Meng artment of atistics, Harvard

Soup

"Trio" Identi

ПP

What's Big?

CCEC

Assessing d.d

Parado

l occon



Effective Sample Size

Menu

8

What's Big?

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

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

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



Effective Sample Size

Menu

8

Department of Statistics.

What's Big?

The Effective Sample Size $n_{\rm eff}$ 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

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



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

Menu

ienu

9

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identii

LLP

What's E

CCES

Assessing d.c

Paradox

essons

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

ienu

9

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identit

T. . .

LLP

What's Big

CCE:

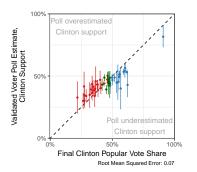
Assessing d.

Paradox

essons

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

/lenu

q

Xiao-Li Meng Department of Statistics, Harvard University

Soul

"Trio" Identit

_.

LLP

What's Big

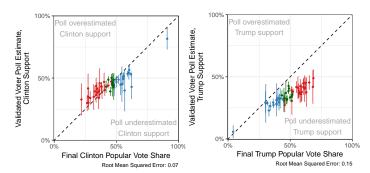
CCE

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

Menu 10

Assessing d.d.i

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

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



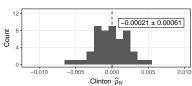
Assessing $\hat{\rho}$ using state-level data

Menu 10

Assessing d.d.i

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

$$\hat{
ho} = rac{\hat{\mu}_N - \mu_N}{\sqrt{rac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N (1 - \mu_N)$$



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



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

10 Menu

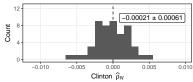
Department of

Assessing d.d.i

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

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

Count





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

-0.0045 ± 0.00056



Menu

.

XIao-LI Mer Department Statistics, Harvard

Soup

"Trio" Identi

Trio

LLP

What's Bi

CCES

Assessing d.d.i

Parado>

esson

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



Menu

Assessing d.d.i

- 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.



Menu

Assessing d.d.i

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

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



Menu

__

Meng nent o

Department of Statistics, Harvard University

Soup

"Trio" Identity

THO TOURING

LLP

What's Big

CCES

Assessing d.d.i

Paradox

lessons

- 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
$$\hat{
ho} = -0.005 = -1/200, D_I = 1/40000$$
, 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 $\hat{\rho} = -0.005$.

Menu

Department of Statistics.

Assessing d.d.i

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
$$\hat{
ho} = -0.005 = -1/200, D_I = 1/40000$$
, 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)

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

Menu

Department of

Assessing d.d.i

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
$$\hat{
ho} = -0.005 = -1/200, D_I = 1/40000$$
, 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)

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

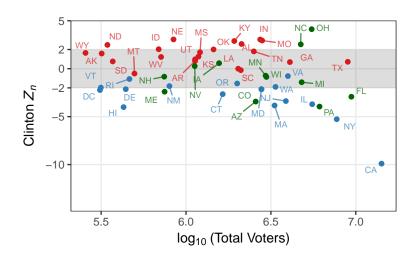
• For N = 230,000,000

Relative Error =
$$-75.8$$



Visualizing LLP: Actual Coverage for Clinton

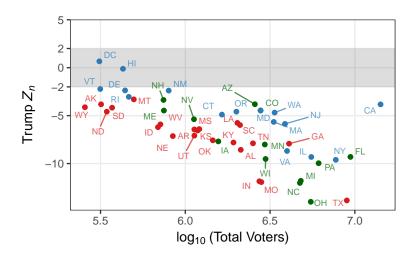
Menu





Visualizing LLP: Actual Coverage for Trump

Menu 13





The Big Data Paradox:

Menu

Department of

Paradox

If we do not pay attention to data quality, then

The bigger the data, the surer we fool ourselves.



Menu

15

Xiao-Li Men Department Statistics,

"Trio" Identit

What's Bi

CCES

Assessing d.d

Parado:

Lessons

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



Menu

15

Xiao-Li Men_i Department o Statistics, Harvard University

Soup

"Trio" Identity

THO Identit

LLP

What's Big

CCEC

Assessing d.d

Paradox

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



Menu

15

Xiao-Li Meng Department of Statistics, Harvard

Soup

"Trio" Identity

Trio

What's Big?

Assessing d.d

Paradox

- 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.



Menu

15

Xiao-Li Meng Department o Statistics, Harvard University

Soup

" Trio" Identit

Trio

What's Big

Assessing d.d.

Paradox

- 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.



Menu

.

Xiao-Li Meng Department of Statistics, Harvard University

Soup

"Trio" Identity

Trio

LLP

What's Big? CCES

Assessing d.d.

- 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 ...



More Lessons From ...

Menu 16

Lessons



A Telescopic, Microscopic, and Kaleidoscopic View of Data Science