

Menu

lenu

Xiao-Li Meng Department o Statistics, Harvard University

Motivation

Soup

Euler Identity

Derivation

_ -----

1110

Whats Big

CCES

Assessing d.d

Parado

essons

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. Annals of Applied Statistics



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. 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

Menu

Motivation

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



Motivating questions

Menu

Department of

Motivation

 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

Menu

Department of Statistics.

Motivation

 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)



Menu

iao-Li Men epartment Statistics, Harvard University

Motivation

Soup

Euler Ident

Derivation

Whats B

CCES

Assessing d.d

Parado

esson

• Law of Large Numbers: Jakob Bernoulli (1713)



Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

Whats Bi

CCES

Assessing d.d

Parado

esson

• Law of Large Numbers: Jakob Bernoulli (1713)

Central Limit Theorem:

Abraham de Moivre (1733):

error $\propto \frac{1}{\sqrt{n}}$: n – sample size



Menu

enu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

ler Identit

Derivation

Whats Big

According d

Parado

esson

• Law of Large Numbers: Jakob Bernoulli (1713)

Central Limit Theorem:

Abraham de Moivre (1733):

error $\propto \frac{1}{\sqrt{n}}$: n – sample size

Survey Sampling:

Graunt (1662); Laplace (1882)



Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sout

Euler Identit

Derivation

Trio

LLP

Whats Big

Assessing d.d

Paradi

esson

• Law of Large Numbers: Jakob Bernoulli (1713)

• Central Limit Theorem: Abraham de Moivre (1733): 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



Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

uler Identit

Derivation

Tric

. . .

Whats Big

CCES

Assessing d.d

Parade

Lesson

- Law of Large Numbers: Jakob Bernoulli (1713)
- Central Limit Theorem: Abraham de Moivre (1733): 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





Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

uler Identit

Derivation

Tric

. . .

Whats Big

CCES

Assessing d.d

Parade

Lesson

- Law of Large Numbers: Jakob Bernoulli (1713)
- Central Limit Theorem: Abraham de Moivre (1733): 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





Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

Trio

LLF

Whats Big

Assessing d.d

Parad

Lessons

• Law of Large Numbers: Jakob Bernoulli (1713)

• Central Limit Theorem: Abraham de Moivre (1733): 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)





Menu

ieiiu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

Euler Identity

Derivation

Trio

LLP

Whats Big

Assessing d.o

Parad

Lessons

- Law of Large Numbers: Jakob Bernoulli (1713)
- Central Limit Theorem: Abraham de Moivre (1733): 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)





Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

Trio

LLP

Whats Big

Assessing d

Parado

Lessons

- Law of Large Numbers: Jakob Bernoulli (1713)
- Central Limit Theorem: Abraham de Moivre (1733): 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





Menu

4

Xiao-Li Mer Department Statistics, Harvard University

Motivation

Soup

Fuler Ident

Derivatio

. . .

Whats B

CCES

Assessing d.d

Parado

esson

Think about tasting soup ...



Menu

1

Xiao-Li Men Department of Statistics, Harvard University

Motivation

Soup

uler Identii

Derivation

Whats B

CCES

Assessing d.c

Paradi

Lesson

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





Menu

ienu

Xiao-Li Meng Department of Statistics, Harvard University

Motivatio

Soup

Euler Identi

Derivation

Whats Bi

CCES

Assessing d.d

Parado

esson

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





Menu

ienu

Xiao-Li Meng Department of Statistics, Harvard University

Motivatio

Soup

Euler Identi

Derivation

Whats E

CCES

Assessing d.d

Parad

Leccono

Think about tasting soup ...

 Stir it well, then a few bits are sufficient regardless of the size of the container!







Menu

lenu

Xiao-Li Meng Department of Statistics, Harvard University

Motivatio

Soup

Euler Identi

Derivation

Trio

. . .

Whats Big

Assessing d.d

Parado

Lessons

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









Menu

Kiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

Euler Identity

Derivation

Trio

Whats Big

CCLS

Assessing d.d

Parado



Menu

Euler Identity



important constants in the subject: 1, 0, m, e and a of Geometry at Gresham College, London, and a former Central to both mathematics and physics, it has also featured in a criminal court case and on a postage stamp, and has appeared twice in The Sympsons. So what is this equation - and why is it pioneering?

Fellow of Keble College, Oxford. 5-6pm Wednesday 28 February 2018 Lecture Theatre 1, Mathematical Institute, Oxford

Mathematics.

Oxford Mathematics Public Lectures

Register at external-relations@maths.ox.ac.uk



Menu

enu

Xiao-Li Meng Department o Statistics, Harvard University

Motivatio

Sou

Euler Identity

Derivatio

. .

Whats Big

Assessing d d

Parado

.

essons



Euler's equation, the 'most beautiful equation in

mathematics', startlingly connects the five most

important constants in the subject: 1, 0, m, e and a

featured in a criminal court case and on a postage stamp, and has appeared being in The Smooths.

So what is this equation - and why is it pioneering?

Mathematics.

Central to both mathematics and physics, it has also

Robin William is an Emeritus Professor of Pure Mathematics at the Open University, Emeritus Professor of Geometry at Gresham College, London, and a former Fellow of Katile College, Oxford.

5-6pm Wednesday 28 February 2018 Lecture Theatre 1, Mathematical Institute, Oxford

Register at external-relations@maths.ox.ac.uk



Menu

Euler Identity

Mathematics.

mathematics', startlingly connects the five most important constants in the subject: 1, 0, m, e and a

Central to both mathematics and physics, it has also

featured in a criminal court case and on a postage

stamp, and has appeared twice in The Simpsons. So what is this equation - and why is it pioneering?



Mathematics at the Open University, Emeritus Professor

of Geometry at Gresham College, London, and a former

Fellow of Keble College, Oxford.

5-6pm Wednesday 28 February 2018

Lecture Theatre 1, Mathematical Institute, Oxford Register at external-relations@maths.ox.ac.uk



 5 most fundamental numbers in mathematics:

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

• The unexpected one:

$$i = \sqrt{-1}$$



Menu

١

ao-Li Meng partment o Statistics, Harvard Jniversity

Motivation

S

Euler Identity

Derivatio

_ -----

Tric

Whats B

CCES

Assessing d.d

Parade

esson

What are the five most fundamental symbols in Statistics?

Average/Mean

$$\mathsf{Ave}\{X_j, j=1,\ldots\}$$



Menu

. '

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soul

Euler Identity

Derivation

116

Whats B

CCES

Assessing d.d

Parad

esson

What are the five most fundamental symbols in Statistics?

• μ : Average/Mean

$$Ave\{X_j, j=1,\ldots\}$$

• σ : Standard Deviation

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$



Menu

,

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

Euler Identity

Derivation

LLI

Whats E

Assessing a.c

Parado

esson

What are the five most fundamental symbols in Statistics?

•
$$\mu$$
: Average/Mean

$$\mathsf{Ave}\{X_j, j=1,\ldots\}$$

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$

$$Ave\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - Ave\left(\frac{X_j}{\sigma_x}\right)Ave\left(\frac{Y_j}{\sigma_y}\right)$$



Menu

Xiao-Li Meng Department of Statistics, Harvard

Motivation

Sou

Euler Identity

Derivation

LLI

vvnats

Assessing d.o

Parade

esson

What are the five most fundamental symbols in Statistics?

•
$$\mu$$
: Average/Mean

$$\mathsf{Ave}\{X_j, j=1,\ldots\}$$

•
$$\sigma$$
: Standard Deviation

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$

$$Ave\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - Ave\left(\frac{X_j}{\sigma_x}\right)Ave\left(\frac{Y_j}{\sigma_y}\right)$$

• n: Sample Size



Menu

Department of Statistics.

Euler Identity

What are the five most fundamental symbols in Statistics?

μ:

Average/Mean Ave
$$\{X_j, j = 1, \ldots\}$$

Standard Deviation \bullet σ :

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$

Correlation ρ:

$$Ave\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - Ave\left(\frac{X_j}{\sigma_x}\right)Ave\left(\frac{Y_j}{\sigma_y}\right)$$

• n: Sample Size

N: Population Size The unexpected one ...



Menu

Department of Statistics.

Euler Identity

What are the five most fundamental symbols in Statistics?

μ:

Average/Mean Ave
$$\{X_j, j = 1, \ldots\}$$

Standard Deviation \bullet σ :

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$

Correlation ρ:

$$Ave\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - Ave\left(\frac{X_j}{\sigma_x}\right)Ave\left(\frac{Y_j}{\sigma_y}\right)$$

• n: Sample Size

N: Population Size The unexpected one ...



Menu

Department of Statistics.

Euler Identity

What are the five most fundamental symbols in Statistics?

•
$$\mu$$
: Average/Mean Ave $\{X_j, j = 1, ...\}$

$$Ave\{X_j, j=1,\ldots\}$$

•
$$\sigma$$
: Standard Deviation $\sqrt{\text{Ave}\{(X_j - \mu)^2\}}$

$$\sqrt{\mathsf{Ave}\{(X_j-\mu)^2\}}$$

$$Ave\left(\frac{X_j}{\sigma_x}\frac{Y_j}{\sigma_y}\right) - Ave\left(\frac{X_j}{\sigma_x}\right)Ave\left(\frac{Y_j}{\sigma_y}\right)$$

• n: Sample Size

N: Population Size The unexpected one ...

The Most Beautiful Statistical Identity?

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



Menu

.

(iao-Li Men epartment Statistics, Harvard

Motivation

auoa

Euler Identi

Derivation

_ -----

Trio

Whats Bi

CCES

Assessing d.d

Parado

eccon

• n: number of respondents to an election survey



Menu

Derivation

• n: number of respondents to an election survey

• N: number of (actual) voters in US



Menu

Derivation

- 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

... I . M...

Department of Statistics, Harvard University

Motivation

Soup

er Identit

Derivation

. . .

Whats Bi

Assessing d.d

Parade

esson

- 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 University

Motivation

Soup

er Identit

Derivation

116

Whats Big

CCES

Assessing d.d

Parado

esson

- 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

Menu

Kiao-Li Men

Department of Statistics, Harvard University

Motivation

Joup

Euler Identity

Derivation

Trio

LLP

CCEC

Assessing d.c

Parado

lesson

- 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:

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



2016 US Presidential Election

Xiao Li Men

Menu

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Euler Identity

Derivation

Trio

Whats Big

Assessing d.d.

Parado

Lessons

- 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
- $R_i = 1$: report (honestly) voting plan; $R_i = 0$ otherwise

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

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

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

8

Xiao-Li Meng Department o Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

Trio

LLF

VVIIats D

CCES

Assessing d.d

Parado

esson

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

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



Data quality, quantity, and uncertainty

Menu

nu

8

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

Euler Identit

Derivation

Trio

LLF

Whats Bi

CCES

Assessing d.d

Parade

Lesson

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

Data Quality

Menu

nu

8

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

Trio

LLF

vviiats Di

Paradi

Lesson

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

Data Quantity

Problem Difficulty



Menu

9

Xiao-Li Meng Department o Statistics, Harvard

Motivation

Soun

Euler Identif

Derivation

Trio

Whats Big

CCES

Assessing d.d

Parado

essons

Mean Squared Error (MSE)

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

Menu

9

Xiao-Li Meng Department of Statistics

Department
Statistics,
Harvard
University

Motivation

Sou

uler Identit

Derivation

Trio

LLE

Whats E

CCES

Assessing d.d

Parade

Lesson

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

Menu

9

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identity

Derivation

Trio

. . .

vvnats b

CCES

Assessing d.d

Parad

Lesson

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

9

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Sou

uler Identit

Derivation

D 0...vat..o.

Trio

LLI

VVhats

A -----1

Danad

lesson

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

Menu 10

LLP

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

Benchmark SRS Standard Error

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



A Law of Large Populations (LLP)

Menu 10

Department of

LLP

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

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

The (lack-of) design effect (Deff)

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

A Law of Large Populations (LLP)

Menu 10

Department of Statistics.

LLP

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

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

The (lack-of) design effect (Deff)

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

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

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

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

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**.

Menu

Whats Big?



Department of Statistics

What's Big? Relative Size or Absolute Size?

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

$$n_{ ext{eff}} = rac{n}{1 + (1 - f)[(N - 1)D_I - 1]} pprox rac{f}{1 - f} rac{1}{\hat{
ho}^2},$$

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

NA-41...41-..

c

Menu

Euler Identit

Danimatian

Delivatio

Iric

Whats Big?

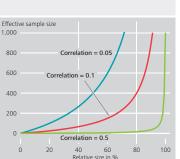
......

Assessing d.d

Parado

.essons

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



• If $\hat{\rho} = 0.05$, then $n_{\rm eff} = 400$ 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

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

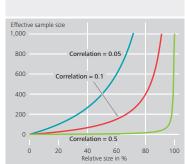
where f = n/N is the **relative size**.

Menu

Department of Statistics.

Whats Big?

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



- If $\hat{\rho} = 0.05$, then $n_{\rm eff} = 400$ 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

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

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

Universi

Department of Statistics

Menu

Motivatio

Эбар

Euler Identii

Derivation

Whats Big?

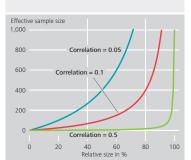
CCEC

Assessing d.d

Parade

essons

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



- If $\hat{\rho}=0.05$, then $n_{\mathrm{eff}}=400$ 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!



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

Menu 1

Xiao-Li Meng Department of Statistics, Harvard

Motivation

Soup

Euler Identi

Derivation

l rı

Whats

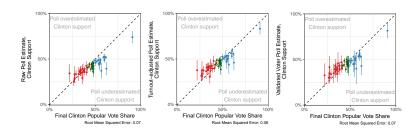
Assessing d.d

Parado

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)



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

Reasonable predictions for Clinton's Vote Share

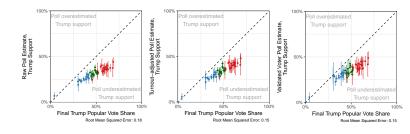


Gross under-prediction/reporting of Trump's Share

Menu

13

CCES: Cooperative Congressional Election Study



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

There are many "undecided" ...



Menu 14

Department o Statistics, Harvard University

Motivation

Soup

Euler Identi

Derivation

Irio

Whats Bi

CCES

Assessing d.d.i

Paradox

raradox



Menu 14

Assessing d.d.i

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

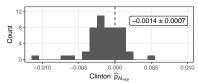
$$\hat{\rho} = \frac{\hat{\mu}_N - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}}, \quad \& \quad \sigma^2 = \mu_N (1 - \mu_N)$$

Menu 14

Assessing d.d.i

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

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



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



Menu 14

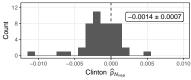
Department of

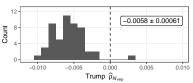
Assessing d.d.i

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

$$\hat{\rho} = \frac{\mu_n - \mu_N}{\sqrt{\frac{N-n}{n}\sigma^2}},$$

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





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



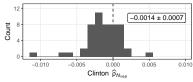
Menu 14

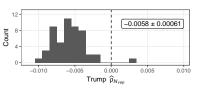
Department of

Assessing d.d.i

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

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





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

Menu

15

Xiao-Li Meng Department of Statistics,

Motivation

Sou

Euler Identit

Derivation

Tric

. . ..

vviiats Dig

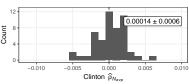
Assessing d.d.i

Parado

_esson

Let $\mu_{\scriptscriptstyle N}$ be the true share, and $\hat{\mu}_{\scriptscriptstyle N}$ 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

Menu

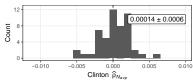
15

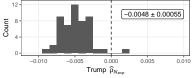
Department of

Assessing d.d.i

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

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





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

Menu

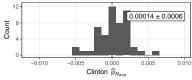
15

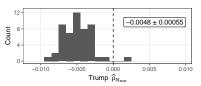
Department of Statistics.

Assessing d.d.i

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

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





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

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

Menu

16

Assessing d.d.i

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

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



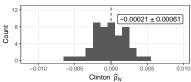
Assessing $\hat{\rho}_{N}$ using validated voter counts

Menu 16

Assessing d.d.i

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

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



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



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

Menu

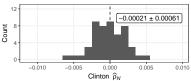
16

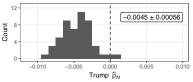
Department of

Assessing d.d.i

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

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





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



Assessing $\hat{\rho}_{\scriptscriptstyle N}$ using validated voter counts

Menu

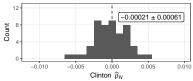
16

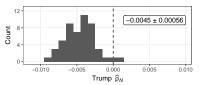
Department of Statistics

Assessing d.d.i

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

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





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.



Menu

Assessing d.d.i

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



Menu

Department of Statistics.

Assessing d.d.i

 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.



Menu

Department of Statistics.

Assessing d.d.i

 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

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

Menu 1

Xiao-Li Meng Department of Statistics, Harvard

Motivatio

Soup

Euler Identit

Derivation

T. . .

LLP

Whats Big

Assessing d.d.i

Parado

Lessor

 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

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

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

Menu 1

Xiao-Li Meng Department of Statistics, Harvard

Motivatio

Soup

Euler Identit

Derivation

T. . .

LLF

Whats Big

Assessing d.d.i

Parado

Lesson

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

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

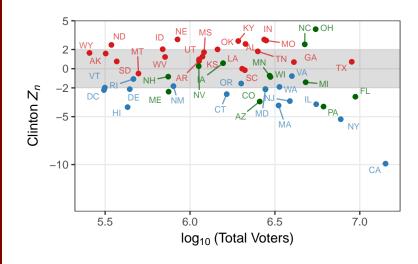
- A 99.98% reduction in n, caused by $\hat{\rho} = -0.005$.
- Butterfly Effect due to Law of Large Populations (LLP)

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



Visulizing LLP: Actual Coverage for Clinton

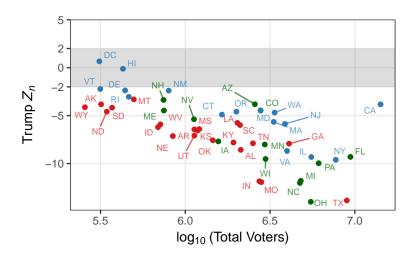
Menu 18





Visulizing LLP: Actual Coverage for Trump

Menu 19





The Big Data Paradox:

Menu

20

Xiao-Li Meng Department of Statistics, Harvard University

Motivation

Soup

uler Identit

Derivation

Whats

CCES

Assessing d.d

Paradox

If we do not pay attention to data quality, then

The bigger the data, the surer we fool ourselves.



Menu

Lessons

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



Menu

Lessons

• 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

Department of

Lessons

- 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

Department of Statistics.

Lessons

- 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

Statistics.

Lessons

 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 ...

Menu

Department of

Lessons

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 eye, 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 =

