ST720 Data Science

Statistical Paradise and Paradoxes in Big Data

Seung Jun Shin (sjshin@korea.ac.kr)

Department of Statistics

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 685–726
https://doi.org/10.1214/18-AOAS1161SF

STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION[1]

BY XIAO-LI MENG

Harvard University

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. By developing measures for data quality, this article suggests a framework to address such a question: "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?" A 5-element Euler-formula-like identity shows that for any dataset of size n, probabilistic or not, the difference between the sample average $\overline{X}_n$ and the population average $\overline{X}_N$ is the product of three terms: (1) a *data quality* measure, $\rho_{R,X}$, the

Introduction I

- Dominating mathematical tool for justifying statistical methods has been large-sample asymptotics.
- Statisticians must be thrilled by the explosive growth of data size.

Introduction II

- However, the reality appears to be the opposite.
  - The size of our data greatly exceeds the volume that can be comfortably handled by our laptops.
  - The variety of the data challenges the most sophisticated models or tools at our disposal.
  - Many problems demand the type of velocity that would make our head spin for both data. processing and analysis.

Introduction III

- The worst is: the more we lament how our nutritious recipes are increasingly being ignored, the more fast food is being produced, consumed and even celebrated as the cuisine of a coming age.
- Some of our most seasoned chefs are working tirelessly to preserve our time-honored culinary skills, while others are preparing themselves for the game of speed cooking.
- Fast food will always exist because of the demand—how many of us have repeatedly had those quick bites that our doctors have repeatedly told us stay away from?

Introduction IV

- Re-inventing the wheel is a well-known phenomenon in almost any field and it is a common source of unhappiness in academia.
- The real damage occurs when the re-invented wheels are inferior, increasing the frequency of serious or even fatal accidents.
- Quality control is thus an important role for statisticians to carry out, as well as a force for innovation because real advances occur more from the desire to improve quality than quantity.

Data Quality-Quantity Tradeoff I

- Main question: Which one should I trust more:

  1% survey with 60% response rate vs non-probabilistic dataset covering 80% of the population

- The qualitative answer clearly is "it depends", on how non-random the larger sample is.

Data Quality-Quantity Tradeoff II

- Analogy of magical power of probabilistic sampling:

  As long as the soup is stirred sufficiently uniformly, a spoonful is all it takes to ascertain the flavor of the soup regardless of the size of its container.

- The quality is measured by the representativeness, achieved via uniform stirring.

Data Quality-Quantity Tradeoff III

- A key question is *how to compare two datasets with different quantities and different qualities?*
- Bigdata NEVER intended to be probabilistic samples

A fundamental Identity I

- For a population, $X_1, \cdots, X_N$, the estimator of $\bar{G}_N$ denoted by $\bar{G}_n$ is

$$\bar{G}_n = \frac{1}{n} \sum_{j \in I_n} G_j = \frac{\sum_{j=1}^{N} R_j G_j}{\sum_{j=1}^{N} R_j} \tag{1}$$

  where $R_j = 1$ for $j \in I_n$ and $R_j = 0$ otherwise.
- $\mathbf{R} = \{R_1, \cdots, R_N\}$ determines the sampling mechanism and we call it R-mechanism.
- For probabilistic random sampling, $\mathbf{R}$ has a well-specified joint distribution.
- Big Data out there, however, they are either self-Reported or administratively recorded.
- Even when the data collector started with a probabilistic sampling design, we have only observations from those who choose to *Respond*.

A fundamental Identity II

- The R-mechanisms are crucial in determining the accuracy of $\bar{G}_n$ as an estimator of $\bar{G}_N$.
- Understand how to quantify the R-mechanisms, and how it affects the accuracy of $\bar{G}_n$.

A fundamental Identity III

- The difference between $\bar{G}_n$ and $\bar{G}_N$ can be written as

$$\begin{aligned}
\bar{G}_n - \bar{G}_N &= \frac{\frac{1}{N}\sum_{j=1}^{N} R_j G_j}{\frac{1}{N}\sum_{j=1}^{N} R_j} - \frac{1}{N}\sum_{j=1}^{N} G_j \\
&= \frac{E_J(R_J G_J)}{E_J(R_J)} - E_J(G_J) \\
&= \frac{E_J(R_J G_J) - E_J(R_J)E_J(G_J)}{E_J(R_J)} \\
&= \frac{\mathrm{Cov}_J(R_J G_J)}{E_J(R_J)}
\end{aligned}$$

where E_J and Cov_J are all taken with respect to the uniform distribution on $J \in \{1, \cdots, N\}$.

A fundamental Identity IV

- Let
  - $\rho_{R,G} = \mathrm{Corr}_J(R_J, G_J)$: Population Correlation between R_J and G_J;
  - $f = E_J(R_J) = \frac{n}{N}$: Sampling Rate;
  - $\sigma_G^2 = \mathrm{Var}_J(G_J)$,

  defined over the uniform distribution of J.
- Notice that $\mathrm{Var}_J(R_J) = f(1 - f)$.

A fundamental Identity V

- Now, we have

$$\bar{G}_n - \bar{G}_N = \underbrace{\rho_{R,G}}_{\text{Data Quality}} \times \underbrace{\sqrt{\frac{1-f}{f}}}_{\text{Data Quantity}} \times \underbrace{\sigma_G}_{\text{Problem Difficulty}} .$$

- Data quality is captured by the data defect correlation $\rho_{R,g}$ because it precisely measures both the sign and degree of selection bias caused by the R-mechanism.
- Under a probabilistic sampling, a particular value of G is recorded/reported or not should not depend on the value itself. (i.e., $E_{\mathbf{R}}(\rho_{R,G}) = 0$)

A fundamental Identity VI

- MSE is given by

$$\begin{aligned} MSE_{\mathbf{R}}(\bar{G}_n) &= E_{\mathbf{R}}[\rho^2_{G,R}] \times \left(\frac{1-f}{f}\right) \times \sigma^2_G \\ &\equiv D_I \times D_O \times D_U \end{aligned}$$

  where $E_{\mathbf{R}}$ denotes the expectation with respect to any chosen distribution of $\mathbf{R}$ given the sample size $\sum_{j=1}^{N} R_j = n$.

- Three ways to reduce MSE:
  - Increase the data quality
  - Increase the data quantity
  - Reduce the difficulty of the estimation problem

Understanding D_I I

- Under SRS, $\bar{G}_n$ is unbiased for $\bar{G}_N$ and its MSE is

$$\mathrm{Var}_{SRS}(\bar{G}_n) = \frac{1-f}{n} S_G^2, \qquad S_G^2 = \frac{N}{N-1}\sigma_G^2$$

which yields

$$D_I \equiv E_{SRS}(\rho_{R,G}^2) = \frac{1}{N-1}$$

A law of large populations I

- Notice that

$$Z_{n,N} \equiv \frac{\bar{G}_n - \bar{G}_N}{\sqrt{\mathrm{Var}_{SRS}(\bar{G}_n)}} = \frac{\rho_{R,G}\sqrt{\frac{1-f}{f}}\sigma_G}{\sqrt{\frac{1-f}{n}S_G^2}} = \sqrt{N-1}\rho_{R,G}$$

Law of Large Populations (LLP)

- When $E_{\mathbf{R}}(\rho_{R,G}) \neq 0$, the (stochastic) error of G_n, relative to its benchmark under SRS, grows with the population size N at the rate of $\sqrt{N}$.

A law of large populations II

- LLP can be expressed in terms of the design effect, or more appropriately the "lack-of-design effect" for non-probabilistic Big Data:

$$
\begin{aligned}
\text{Deff} &= E_{\mathbf{R}}(Z_{n,N}^2) \\
&= \frac{E_{\mathbf{R}}[\bar{G}_n - \bar{G}_N]^2}{\mathrm{Var}_{SRS}(\bar{G}_n)} \\
&= (N-1)E_{\mathbf{R}}(\rho_{\rho,G}^2) = (N-1)D_I.
\end{aligned}
$$

A law of large populations III

Theorem

- For a fixed sampling rate $0 < f < 1$ and problem difficulty $D_U = \sigma_G^2$, the following three conditions are equivalent for any R-mechanism:
  1. It has a finite design effect: Deff $= O(1)$;
  2. The MSE of the sample mean decreases at the n^{-1} rate: $MSE_{\mathbf{R}}(\bar{G}_n) = O(n^{-1})$.
  3. Its d.d.i for the sample mean is controlled at the N^{-1} level: $D_I = O(N^{-1})$.

A law of large populations IV

- For large populations achieving $\rho_{R,G} \approx N^{-1/2}$ without probabilistic sampling requires a miracle.
- 2016 US population with $N \approx 1.4 \times 10^8$. We require

$$\rho_{R,G} \approx N^{-1/2} = 8.4 \times 10^{-5},$$

an extremely small correlation coefficient to be guaranteed from a self-regulated selection mechanism.

A butterfly effect: The return of the long-forgotten monster N I

- To quantify the damage by $\rho_{R,G}$, let's use the effective sample size n_{eff} of a Big Data by equating the MSE of its estimator $\bar{G}_n$ to that of the SRS estimator with the sample size n_{eff}.
- Recall that

$$MSE_{\mathbf{R}}(\bar{G}_n) = \frac{1}{N-1} \frac{1 - f_{\text{eff}}}{f_{\text{eff}}} \sigma_G^2 = D_I D_O D_U$$

which yields

$$D_I D_O = \left(\frac{1}{n_{\text{eff}}} - \frac{1}{N} \right) \frac{N}{N-1}$$

A butterfly effect: The return of the long-forgotten monster N II

- Thus we have

$$n_{\text{eff}} = \frac{n^*_{\text{eff}}}{1 + (n^*_{\text{eff}} - 1)N^{-1}}$$

where $n^*_{\text{eff}} = (D_I D_O)^{-1}$.

- Under the (trivial) assumption that $n^*_{\text{eff}} \geq 1$, we have

$$n_{\text{eff}} \leq n^*_{\text{eff}} = \frac{f}{1-f} \times \frac{1}{D_I} = \frac{n}{1-f} \times \frac{1}{ND_I}$$

  - For probabilistic sample, N is canceled out by $D_I = O(N^{-1})$.
  - When $D_I = O(1)$, however small, ND_I increases with N quickly, leading to a dramatic reduction of n_{eff}.

A butterfly effect: The return of the long-forgotten monster N III

- Suppose $E_{\mathbf{R}}[\rho_{R,G}] = 0.05$,

$$D_I = E_{\mathbf{R}}(\rho^2) \geq [E_{\mathbf{R}}(\rho_{R,G})]^2 = \frac{1}{400}$$

  which yields

$$n_{\text{eff}} \leq 400 \frac{f}{1-f}$$

- When $f = 0.5$, the effective sample size, in terms of an equivalent SRS sample, cannot exceed $n_{\text{eff}} = 400$.
- For the US population in 2016, we have

$$\frac{115,000,000 - 400}{115,000,000} = 99.999965\%$$

  reduction of the sample size.

A butterfly effect: The return of the long-forgotten monster N IV

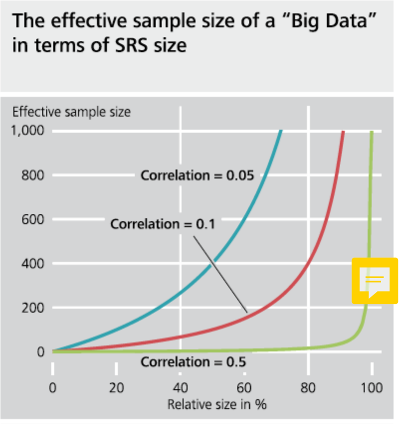

Figure: n^*_{eff} as a function of f for different values of $E_{\mathbf{R}}[\rho_R, G]$.

A butterfly effect: The return of the long-forgotten monster N V

- Recall that $n^*_{\text{eff}} = (D_I D_O)^{-1}$.
- Even if n is practically large, $D_O = \frac{1-f}{f}$ is not very close to 0 when N is really large.
- Once we lose control over the R mechanism via probabilistic schemes, the design effect $(N-1)D_I$ explodes.
- We have a "butterfly effect"-a tiny perturbation caused by D_I can lead to catastrophic error in the end for large N.

A big data paradox? I

- Consider the following CI

$$\left(\bar{G}_n - \frac{M\hat{\sigma}_G}{\sqrt{n}}, \bar{G}_n + \frac{M\hat{\sigma}_G}{\sqrt{n}}\right)$$

  for a given constant M.
- By LLP, as N and n gets larger while $f < 1$ fixed, we almost surely miss $\bar{G}_N$ for any M.
- Moreover, the MOE shrinks toward to 0 as n increases.

Bigdata Paradox

- The bigger the data, the surer we fool ourselves.

Answering the motivating question I

- Our first dataset is a probabilistic sample with sampling rate $f_s = n_s/N$ and design effect "Deff".
- Without non-response, we know $D_I^{(s)} = \text{Deff}/(N-1)$.
- With non-response, we know that D_I is larger than $D_I^{(s)}$, and $D_O = \frac{1-rf_s}{rf_s}$ is larger than $D_O^{(s)} = \frac{1-f_s}{f_s}$.

Answering the motivating question II

- Second data is a Big data with D_I^{BIG} and D_O^{BIG}.
- Then $n_{\text{eff}}^{\text{BIG}}$ is larger than the n_{eff} of the 1st data set iff

$$D_I^{\text{BIG}} D_O^{\text{BIG}} < D_I D_O$$

- Equivalently

$$|\rho_{R,G}^{\text{BIG}}| \leq \sqrt{\mathcal{O}} |\rho_{R,G}|$$

where the dropout odds ratio $\mathcal{O}$ is given by

$$\mathcal{O} = \frac{D_O}{D_O^{\text{BIG}}} = \frac{1 - rf_s}{rf_s} \frac{f}{1-f}.$$

Answering the motivating question III

- For our question,

$$f_s = 0.01, r = 0.6 \ \text{ and } \ f = 0.8$$

which yields

$$\sqrt{\mathcal{O}} \approx 26$$

- If we are reasonably sure that the mechanism leading to non-response in our survey is similar to the mechanism responsible for self-reporting behavior in the Big Data, then we should be reasonably confident that the Big Data set is more trustworthy.

Answering the motivating question IV

- If we believe that the selection bias caused by the nonresponse mechanism in the sample is not nearly as severe as in the Big Data set.
- We need to have a reasonable sense of the magnitude of $\rho_{R,G}$. The population size is useful for this assessment.
- For $N \approx 231,557,000$, we have

$$|\rho^{(s)}| \approx \sqrt{2/\pi}(N-1)^{-1/2} = 5.2 \times 10^{-5}$$

since $E|Z| = \sqrt{2/\pi}$.

Answering the motivating question V

- Suppose the non-response mechanism has increased the data defect correlation 5 times, then

$$\sqrt{\mathcal{O}} \times \rho_{R,G} \approx 26 \times 2.6 \times 10^{-4} = 0.0068.$$

- For trump's supporters in the US election 2016, $\rho_{R,G}^{\mathsf{BIG}}$ is assessed as -0.005, thus the Big data is still more trustworthy.
- When $f = 0.5$ (50% of population) instead of 0.8, $\sqrt{\mathcal{O}} \approx 13$ and thus $|\rho_{R,G}^{\mathsf{BIG}}| < 0.0034$ in order to trust the Big Data. Therefore, the first data set is more trustworthy.