Statistical Paradises and Paradoxes in Big Data (I):

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

By Xiao-Li Meng

Summary by Thea Quistgaard

"The bigger the data, the surer we fool ourselves"

March 7, 2019

- Sample vs. population
- **Probabilistic sampling:** Each subject has some given probability and the sample is drawn given this distribution. E.g. Simple Random Sampling (SRS)
- **Non-probabilistic sampling:** based on the subjective judgment of the researcher rather than random selection. Not all subjects have probability of being drawn. E.g. Election polls

An Interesting Question...

"*Which one should I trust more: a 1% survey with 60 % response rate or a non-probabilistic dataset covering 80 % of the population?*"

- Data quality
- Data quantity
- Problem difficulty

CAN WE SOMEHOW LINK THESE IDENTITIES?

- Data quality
- Data quantity
- Problem difficulty

CAN WE SOMEHOW LINK THESE IDENTITIES?

Well, yes, of course...

$$\bar{G}_n - \bar{G}_N = \rho_{R,G} \cdot \sqrt{\frac{1-f}{f}} \cdot \sigma_G \tag{1}$$

- **Data Quantity Measure**: $\sqrt{\frac{1-f}{f}}$ ($f = \frac{n}{N}$, relative sample size)
- **Problem Difficulty**: σ_G, the variation over G
- **Data Quality Measure**: $\rho_{R,G}$, *data defect correlation* with $R_J = 1$ if $j \in$ sample: recording/response mechanism

$$\begin{aligned} MSE_R(\bar{G}_n) &= E_R[\bar{G}_n - \bar{G}_N]^2 \\ &= E_R[\rho^2_{R,G}] \cdot \frac{1-f}{f} \cdot \sigma^2_G \\ &\equiv D_I \cdot D_O \cdot D_U \end{aligned} \tag{2}$$

- **Increase data quality** by reducing $D_I = E_R[\rho^2_{R,G}]$ the *Data Defect Index (d.d.i.)*.
- **Increase the data quantity** by reducing the Dropout Odds, $D_O = \frac{1-f}{f}$.
- **Reduce the difficulty** of the problem by reducing the Degree of Uncertainty, $D_U = \sigma^2_G$.

(1) "*What are the likely magnitudes of* D_I *when we have probabilistic samples?*"

- $V_{SRS}(\bar{G}_n) = \frac{1-f}{n} \frac{N}{N-1} \sigma_G^2$
- $D_I \equiv E_{SRS}[\rho_{R,G}^2] = \frac{1}{N-1}$
- $D_I \propto N^{-1}$ *holds in general for any probabilistic sampling*

(2) "*How do we calculate or estimate D_I for non-probabilistic data?*"

- Not possible to estimate from sample itself
- Construct a reasonable prior distribution of $\rho_{R,G}$ from historical or neighboring studies.

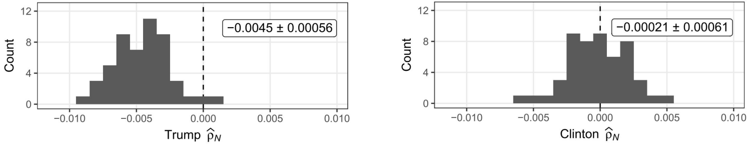

Figure: Trump and Clinton polls (1)

PROBABILISTIC: A usual driving force for stochastic behaviors is the sample size n.

- Central Limit Theorem
- Law of Large Numbers

NON-PROBABILISTIC: The driving force is actually the *population size, N*.

$$
\begin{aligned}
Z_{n,N} &\equiv \frac{\bar{G}_n - \bar{G}_N}{\sqrt{V_{SRS}}} \\
&= \frac{\rho_{R,G}\sqrt{\frac{1-f}{f}}\sigma_G}{\sqrt{\frac{1-f}{n}\frac{N}{N-1}\sigma_G^2}} \\
&= \sqrt{N-1}\rho_{R,G}
\end{aligned} \tag{3}
$$

Among studies sharing the same (fixed) average data defect correlation $E_R[\rho_{R,G}] \neq 0$, the stochastic error of $\bar{G}_n$, relative to its benchmark under SRS, grows with population size N at the rate of $\sqrt{N}$.

The effective sample size

$$n_{eff} \leq \frac{f}{1-f} \cdot \frac{1}{D_I}. \tag{4}$$

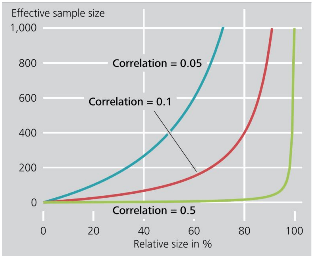

Figure: Illustration of n_{eff} compared to the relative size.[1]

[1] Figure from Mehrhoof (2016)(2)

- Sometimes quality over quantity
- Beware of your recording/response mechanisms
- *The more the data, the surer we fool ourselves.*

[1] MENG, X.-L. (2018), Harvard University *Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election*, The Annals of Applied Statistics, 2018, Vol. 12, No 2, 685-726.

[2] MEHRHOFF, J. (2016). Executive summary: Meng, X.-L. (2014), "A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)". Conference handout.

[3] MCDONALD, M. P. (2017). 2016 November general election turnout rates.