The biggest enemy to reliable evidence from Data Science:

Selection Bias

(intentional or unintentional)

Xiao-Li Meng

Founding Editor-in-Chief

HARVARD DATA SCIENCE REVIEW

Everything Data Science and Data Science for Everyone

Cherry picking alters the strength of evidence

- Lack of evidence
- Preponderance of evidence
- Beyond a reasonable doubt

$$\text{Probability (or risk)} = \frac{\#\ \text{Cases of Interest}}{\text{Total}\ \#\ \text{of Cases}} = \frac{n}{N}$$

Cherry picking almost always alters N, and sometimes n as well, and it aims to drive the ratio to 0 or 1.

https://www.ariasolutions.com/cherry-picking-work-is-bad/

7 S’(ins) even if we don’t intend to cherry pick

1. Selections in target/hypotheses (e.g., subgroup analysis)
2. Selections in data (e.g., deleting “outliers” or using only “complete cases”)
3. Selections in methodologies (e.g., for goodness of fit)
4. Selections in due diligence and debugging (e.g., triple checking only when the outcome seems undesirable)
5. Selections in publication (e.g., only when p-value <0.05)
6. Selections in reporting/summary (e.g., suppressing caveats)
7. Selections in understanding and interpretation (e.g., our preference for deterministic,“common sense” interpretation)

KEY PROBLEM: Any selection process changes N or n, yet we don’t (know how to) quantify the change.

The devastating impact of selection bias on estimating COVID positive rate

- Selection correlation $\boldsymbol{\rho}$: high risk people are more likely get tested
- f = testing rate/sampling rate

$$\textbf{Effective Sample Size} \cong \frac{f}{1-f} \times \frac{1}{\rho^2}$$

- NY State: N=19.4 M, suppose we conduct ***n*=10,000 tests** (f =1/2000) and the selection effect is a ½ **percent correlation** ($\boldsymbol{\rho}$=0.005):

Same as conducting $\dfrac{0.0005}{0.9995} \times \dfrac{1}{0.005^2} \approx$ **20 random tests!**

- **A 99.80% loss of sample size due to selection bias**

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 Vol 2: 685-726*

Selection bias is inevitable, because it works.

By asking the right questions, we can be less fooled.

- When was this study published? How many related studies did you go though? How many of them reached similar conclusions as the one you reported?
- How many cases were collected? How many of them were used in this study?
- Who collected the data? Who cleaned them? Were any data discarded? Why?
- To which reference population is this case compared? How was this reference population chosen? Why is it so large/small? What happens if we change the population to …?
- ….

What should we look out/for ?

2021/3/17

- Real Estate:

  Location, Location, Location

- Data Science:

  Selection, Selection, Selection

"If you torture the data long enough, it will confess to anything."

Ronald Coase

https://unsplash.com/s/photos/