Multi-source Inference: Surprises and Challenges

Xiao-Li Meng*
Harvard University, Cambridge, MA 02138, U.S.A.
meng@stat.harvard.edu

Statistical inference is a field full of problems whose solutions need the same intellectual force necessary for winning a Nobel Prize in other scientific fields. One of the most recent ones is *multi-source inference*, which aims to extract desired information in data coming from very different sources, some of which were never collected for statistical analysis purposes, such as governmental administrative records (e.g., employment insurances records, tax records, etc.). These non-statistical datasets tend to cover a very large percentage of a population (e.g., over 90%), a fact that has been used by many as a reason for overlooking selection biases inherent in these non-statistical data. But is 90% non-statistical dataset safer than, say, a 0.9% statistical sample? How should we combine data sources with very different statistical qualities? Actually, how should we quantify data qualities? This talk reveals findings that may surprise some, disturb others, and ultimately challenge all of us to think deeper and work harder in order to ensure the most important V of dealing with Big Data: Validity, going beyond the well-known triple-V (Volume, Velocity, and Varity) characterizations of Big Data.

Keywords: Administrative Records; Big Data; Data Defect Index; Selection Bias.