

**(Data) Size Does Matter,  
But You Might Be In for a Surprise ...**

Xiao-Li Meng

*Department of Statistics, Harvard University*

One of the most frequently asked questions in statistical practice, and indeed in general quantitative investigations, is “What is the size of the data?” (I.e., how much information there is in the data?) A common wisdom underlying this question is that the larger the size, the more trustworthy are the results. Although this common wisdom serves well in many practical situations, sometimes it can be devastatingly deceptive. This talk will report two of such situations: a historical epidemic study (McKendrick, 1926) and the most recent debate over the validity of multiple imputation inference for handling incomplete data (Meng and Romero, 2003). McKendrick's mysterious and ingenious analysis of an epidemic of cholera in an Indian village provides an excellent example of how an apparently large sample study (e.g.,  $n=223$ ), under a naive but common approach, turned out to be a much smaller one (e.g.,  $n<40$ ) because of hidden data contamination. The debate on multiple imputation reveals the importance of the self-efficiency assumption (Meng, 1994) in the context of incomplete-data analysis. This assumption excludes estimation procedures that can produce more efficient results with less data than with more data. Such procedures may sound paradoxical, but they indeed exist even in common practice. For example, the least-squared regression estimator may not be self-efficient when the variances of the observations are not constant. The morale of this talk is that in order for the common wisdom “the larger the better” be trusted, we not only need to assume that data analyst knows what s/he is doing (i.e., an approximately correct analysis), but more importantly that s/he is performing an efficient, or at least self-efficient, analysis.