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)," in: Lin, X. et al. (Eds.), *Past, Present, and Future of Statistical Science*, CRC Press, 537-562.

"*Is an 80% non-random sample 'better' than a 5% random sample in measurable terms? 90%? 95%? 99%?*" (Wu, 2012)

Large absolute size or large relative size?

Let us consider a case where we have an administrative record covering f_a percent of the population, and a simple random sample (SRS) from the same population which only covers f_s percent, where $f_s \ll f_a$. How large should f_a/f_s be before an estimator from the administrative record dominates the corresponding one from the SRS, say in terms of MSE?

As an initial investigation, let us denote our finite population by $\{x_1, \dots, x_N\}$. For the administrative record, we let $R_i = 1$ whenever x_i is recorded and zero otherwise; and for SRS, we let $I_i = 1$ if x_i is sampled, and zero otherwise, $i = 1, \dots, N$. Here we assume $n_a = \sum_{i=1}^{N} R_i \gg n_s = \sum_{i=1}^{N} I_i$, and both are considered fixed in the calculations below. Our key interest here is to compare the MSEs of two estimators of the finite-sample population mean $\bar{X}_N$, namely,

$$\bar{x}_a = \frac{1}{n_a}\sum_{i=1}^{N} x_i R_i \text{ and } \bar{x}_s = \frac{1}{n_s}\sum_{i=1}^{N} x_i I_i.$$

Recall for finite-population calculations, all x_i's are fixed, and all the randomness comes from the response indicator R_i for $\bar{x}_a$ and the sampling indicator I_i for $\bar{x}_s$. The administrative record has no probabilistic mechanism imposed by the data collector.

Expressing the exact error, where $f_a = n_a/N$.

$$\bar{x}_a - \bar{X}_N = \frac{\mathrm{E}[xR]}{\mathrm{E}[R]} - \mathrm{E}[x] = \frac{\mathrm{Cov}[x,R]}{\mathrm{E}[R]} = \underbrace{\rho_{x,R}}_{\text{Data Quality}} \cdot \underbrace{\sigma_x}_{\text{Problem Difficulty}} \cdot \underbrace{\sqrt{\frac{1-f_a}{f_a}}}_{\text{Data Quantity}}.$$

Given that $\bar{x}_s$ is unbiased, its MSE is the same as its variance.

$$\mathrm{Var}[\bar{x}_s] = \frac{1-f_s}{n_s} S_N^2(x), \text{ where } S_N^2(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x}_N)^2.$$

The MSE of $\bar{x}_a$ is more complicated, mostly because R_i depends on x_i.

$$\mathrm{MSE}[\bar{x}_a] = \mathrm{E}\left[\rho_{x,R}^2\right] \cdot \sigma_x^2 \cdot \left(\frac{1-f_a}{f_a}\right).$$

It is worthy to point out that the seemingly mismatched units in comparing the relative size f_a with the absolute size n_s reflect the different natures of non-sampling and sampling errors. The former can be made arbitrarily small only when the relative size f_a is made arbitrarily large, that is $f_a \to 1$; just making the absolute size n_a large will not do the trick. For biased estimators resulting from a large self-selected sample, the MSE is dominated (and bounded below) by the squared bias term, which is controlled by the relative sample size.

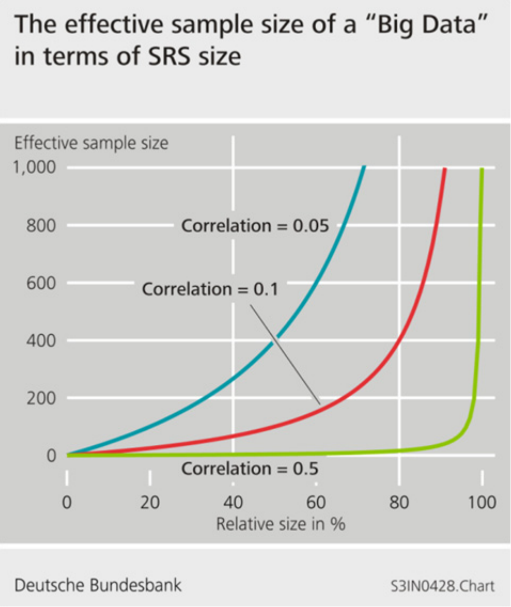
To guarantee $\mathrm{MSE}[\bar{x}_a] \leq \mathrm{Var}[\bar{x}_s]$, we must require (ignoring the finite population correction $1 - f_s$)

$$f_a \geq \frac{n_s \rho_{x,R}^2}{1 + n_s \rho_{x,R}^2}, \text{ or equivalently } n_s \leq \left(\frac{f_a}{1 - f_a}\right) \frac{1}{\rho_{x,R}^2} = \left(\frac{n_a}{N - n_a}\right) \rho_{x,R}^{-2}.$$

We must be mindful, however, that these comparisons assume the SRS and more generally the survey data have been collected perfectly, which will not be the case in reality because of both non-responses and response biases (the SRS will also have a non-zero $\rho_{x,I}$). Hence in reality it would take a smaller f_a to dominate the probabilistic sample with f_s sampling fraction, precisely because the latter has been contaminated by non-probabilistic selection errors as well. Nevertheless, a key message here is that, as far as statistical inference goes, what makes a “Big Data” set big is typically not its absolute size, but its relative size to its population. Therefore, the question which data set one should trust more is unanswerable without knowing N. But the general message is the same: when dealing with self-reported data sets, do not be fooled by their apparent large sizes or by common wisdom from studying probabilistic samples.

Data defect index

The re-expression of the bias in terms of the correlation between sampling variable x and response indicator R is a standard strategy in the survey literature. Although mathematically trivial, it provides a greater statistical insight, that is, the sample mean from an arbitrary sample is an unbiased estimator for the target population mean if and only if the sampling variable and the data collection mechanism are uncorrelated. In this sense we can view $\rho_{x,R}$ as a “defect index” for estimation (using sample mean) due to the defect in data collection/responding. Of course all these calculations depend critically on knowing the value of $\rho_{x,R}$, which cannot be estimated from the biased sample itself.



The effective sample size of a “Big Data” in terms of SRS size

Deutsche Bundesbank S3IN0428.Chart

Imagine that we are given a SRS with $n_s = 400$. If $\rho_{x,R} = 0.05$ and our intended population is the USA, then $N \approx 320{,}000{,}000$, and hence we will need $f_a = 50\%$ or $n_a \approx 160{,}000{,}000$ to place more trust in $\bar{x}_a$ than in $\bar{x}_s$. If $\rho_{x,R} = 0.1$, we will need $f_a = 80\%$ or $n_a \approx 256{,}000{,}000$ to dominate $n_s = 400$. If $\rho_{x,R} = 0.5$, we will need over 99% of the population to beat a SRS with $n_s = 400$.

This reconfirms the power of probabilistic sampling and reminds us of the danger in blindly trusting that “Big Data” must give us better answers. Lesson learned: What matters most is the quality, not the quantity.

This summary is also based on an extremely timely paper presented by X.L. Meng at the RSS 2016 Annual Conference, “Statistical Paradises and Paradoxes in Big Data”. E-mail: jens.mehrhoff@bundesbank.de. All remaining errors are my sole responsibility.