Additional Details on the Table 2 Simulations in "Beyond the Existence Proof"

Samuel R. Lucas

The simulations in Table 2 compare the correlations obtained via probability sampling and those obtained via non-probability sampling the same population. The motivation for the simulation is that some non-probability samplers claim that probability sampling is only useful for calculating means. Such claims imply that non-probability sampling will be more effective at reflecting relationships. The correlation coefficient is one measure of the relationship between two variables. Thus, assessing the performance of probability and non-probability samples in obtaining the correlation coefficient is to assess their performance in reflecting relationships.

I constructed a population of 1,000,000 cases, with measures on 4 variables. I set the correlations specifically, and confirmed the exact values empirically.

I then drew ten probability samples, obtaining the correlation matrix for each sample. This was straightforward, and the effort can be replicated using the bepprobsamp.inc file.

Obtaining a non-probability sample was more complicated. First, people are more similar to those in their network than they are to a person drawn at random. Reversing the reasoning this observation implies, I constructed a measure of the proximity of cases to each other. I then ordered the cases by this measure, such that the cases closest to a given case were those in their network.

For each of ten iterations I then drew a probability sample of 1. Using this one case, I then drew in 39 other members of their network.

Two observations are in order. First, the use of a probability sample of 1 to start the process reflects that there was no way to really obtain a *fully* non-probability sample because the starting point, at some point in the simulation, *must* be random. There is no way to approximate a non-random set of starting points. The impact of this on results is probably to make them more able to capture population characteristics.

Second, I used only 1 (random) starting point. I did so because using more than one would exacerbate the problem above, of moving further away from a fully non-probability sample. Of course, multiple starting points would more closely approximate the way snowball samplers proceed. Yet, their multiple starting points may have much less value than presumed. Because multiple snowball sampling starting points likely share the same unknown biases (e.g., accessing the more accessible members of any network), using one starting point in the simulation accurately reflects this limitation of the method.

Analysts can access the bepnonprobsamp.inc file to run additional simulations.

Use of bepprobsamp.inc or bepnonprobsamp.inc, either as is or as a starting point for other work, should be acknowledged in presented or published work, and a citation to "Beyond the Existence Proof" should be included.