Survey Weighting with Differentially Private Releases of Population Data: An Evaluation of the Cooperative Congressional Election Study

Bhaven Patel and Anthony Rentsch

One of the key methodological components of survey research is weighting, a technique that tries to correct imbalances between the composition of the population sample and the composition of the true population of interest. A typical example is sampling American adults. For a variety of reasons, it is next to impossible to collect a simple random sample of the American adult population, so survey researchers must gather samples through other methods, which often lead to samples that do not accurately reflect the American adult population. For instance, surveys conducted via landline telephones tend to overrepresent older Americans and underrepresent younger Americans. Staying with this example, survey researchers would compute weights that give more importance to responses from younger Americans and less importance to responses from older Americans.

There are numerous different methods to compute survey weights. For more, see this brief summary compiled by the Pew Research Center[1]. There is one thing in common across these methods, though: the reliance on American Census Bureau data to estimate true population parameters. In particular, the Census' American Community Survey (ACS) is a high-quality survey that is typically regarded as a reliable population benchmark.

Beginning in 2020, the Census will release all of its data products in a differentially private manner. Crucially, this includes the ACS. While the exact approach the Census will pursue to release ACS data remains unclear, there is a possibility weighting surveys to ACS data released in a differentially private manner will suffer from new issues not encountered in previous ACS releases. To our knowledge, this is not a topic area that has received attention. One recent study looked at generating survey weighted frequency tables under differential privacy, but their analysis focused more on the effects of adding noise to survey data that had already been weighted rather than adding noise to the benchmark data which is weighted to[2].

Our project will look at the possible effects of differentially private ACS releases on the weighting for one survey in particular: the Cooperative Congressional Election Study (CCES). The CCES is a large-N national survey conducted in every election year. The survey employs a fairly complex sampling and weighting procedure that consists of (1) constructing a sampling frame via a blend of ACS, Current Population Study (CPS), and the Pew Religious Landscape Survey benchmarks; (2) stratifying their panel using these benchmarks; (3) conducting simple

[1] https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/

[2] Shlomo, Krenzke, and Li. 2018. *Comparison of Post-tabular Confidentiality Approaches for Survey Weighted Frequency Tables*.

random sampling from within each strata; (4) matching the sample to a target frame; and (5) using entropy balancing to weight the matched sample to the sampling frame[3].

Our idea to study the effects of the Census' changes in data release practices on the CCES' survey weights is two-fold: (1) we will backtest the weights obtained for several previous CCES surveys by altering the ACS releases from those years to comply with differential privacy, but still assuming that the noisy releases are tried as the point estimates of the ground truth, and (2) we will evaluate different methods to explicitly model the differentially private mechanism in the weighting procedure.

Combining data from the ACS, CPS, and Pew Religious Landscape Survey will be difficult since the CCES does not explicitly state how data from the three sources is coalesced into the sampling frame. We do worry that replicating this weighting procedure will be too difficult and are actively seeking other datasets with simpler weighting procedures and advice from prior colleagues (who work in this area) for this project. If our efforts to replicate the CCES weighting procedure or find another source dataset are not fruitful, we propose generating sample weights for the CCES data using just the population counts derived ACS data, since both datasets are available to us. To calculate these weights, we will use raking, which is the most prevalent method for weighting[1]. Using actual ACS data, we will generate sample weights, which we will compare to sample weights derived from differentially private released ACS data. We would like to test a few differentially private methods, such as the Laplace mechanism and differentially private histograms, for releasing the ACS data and determining their utility when compared to the normal sample weights from the true ACS data. We will also test various values of epsilon, the privacy-loss parameter, to determine a good trade-off between privacy and utility.

Additionally, we are interested in understanding how our sample weights derived from differentially private released ACS data will affect the reporting of actual survey results from the CCES. We would like to see how estimates for political party vote share on a state by state basis, presidential approval and support for certain policies, such as gun control and taxes, vary when normal weights are used and when our DP weights are used.

Timeline of work

- 4/9 - 4/14: Follow up with contacts regarding datasets and advice. Collect ACS and CCES data and understand the codebook for each.
- 4/15 - 4/21: Derive sample weights using raking for CCES data based on true ACS data. Begin developing different methods for differentially private releases of ACS data.

[3] This is the methodology used in the 2016 CCES. The methodology changes slightly year-to-year.

- 4/22 - 4/28: Apply raking to differentially private released ACS data. Compare normal sample weights to sample weights from differentially private released data. What does the utility-privacy tradeoff look like?
- 4/29 - 5/5: Determine effect of differentially private released weights on survey results. Can we improve results if needed.
- 5/6 - final presentation date: Finish any lingering tasks and write-up our results for our presentation and paper.