Survey Weighting with Differentially Private Releases of Population Data

An Evaluation of the Cooperative Congressional Election Study

Bhaven Patel and Anthony Rentsch

May 10, 2019

Abstract

Beginning in 2020, the American Census Bureau has announced that it will release its data products in a differentially private (DP) fashion, a change which introduces new considerations for researchers who rely on Census data. In this paper we focus on the effects of differential privacy on statistical methods that use Census data for calibration; in particular, we look at the case of survey weighting. We find that, for smaller populations, weighting a survey to a differentially private release of the American Community Survey can shift estimates of important survey items by a small margin, but that this shift vanishes when the privacy loss parameter is incremented, i.e., when less noise is added to the released data.

1. Introduction

A key methodological component of survey research is weighting, a technique that tries to correct imbalances between the composition of the sample and the composition of the true population of interest. For a variety of reasons, it is extremely difficult 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. To obtain credible population-level estimates, it is necessary to re-weight individual respondents in a sample to accurately reflect the population as a whole.

While there are numerous different methods to compute survey weights, there is one common feature across these methods: the reliance on American Census Bureau data to estimate true, benchmark population parameters. In particular, the Census' American Community Survey (ACS) is a high-quality survey that produces different estimates of population demographics and socioeconomic conditions in specific geographic regions in the United States, and is typically regarded as a reliable population benchmark.

Beginning in 2020, the Census has announced that it will release all of its data products in a differentially private (DP) fashion in order to protect the confidentiality of individuals who participate in the decennial census, as well as in other surveys that the bureau fields. Differential privacy is a mathematical formulation of privacy that provides guarantees about what an adversary is able to learn about any individual in a dataset after performing some statistical analysis to glean information from the dataset. Intuitively, differential privacy adds noise to a statistical or machine learning analysis so that it is difficult to tell whether or not any individual's information was used as a part of the analysis.

Crucially, the Census Bureau's plan includes the ACS data. The exact approach the Census will pursue to release any of its data products remains unclear, and research is currently being conducted regarding the mechanisms that will be employed. Regardless of what the final implementation is, it is quite likely that the work of analysts who use Census data will be impacted in a manner that has not been encountered before with previous Census data releases [1].

The focus of this paper is on statistical models that rely on Census data for calibration. Specifically, we will analyze the impact that a differentially private release of the ACS could have on the computation of survey weights for one highly-used, large sample size political survey: the Congressional Cooperative Election Study (CCES). We compare subgroup estimates for various survey items when the survey is weighted to the actual ACS data versus when the survey is weighted to a hypothetical differentially private release of the ACS data. We find that, for smaller populations, weighting the CCES to a differentially private ACS release can shift estimates of important survey items by a small margin, but that this shift vanishes when the privacy-loss parameter is incremented, i.e., when less noise is added to the released data.

To our knowledge, this is not a topic area that has received scholarly 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 to which the survey is weighted [3].

We emphasize that this analysis is a first cut at understanding the impact of differential privacy on statistical calibration methods for two reasons: (1) there still remains an incredible amount of uncertainty regarding how the Census Bureau will implement differential privacy for the ACS and (2) surveys like the CCES employ incredibly complex sampling and weighting procedures, which are beyond the scope of this project to replicate. Future work will need to take into account both of these considerations more seriously. For now, we think that our stylized example provides a useful look into what new issues consumers of Census data will face with the rise of differential privacy.

2. Key Background

2.1 Differential privacy

Differential privacy is a criteria for a statistical or machine learning mechanism that holds that an adversary should be able to make the same inference about an individual's sensitive information whether or not that individual was counted in the analysis [4]. In mathematical notation, this implies that $P[M(x) \in T] \leq e^{\epsilon} P[M(x') \in T]$ for some mechanism M, two neighboring datasets x and x' differing on just one row, and some set T. In this definition, ϵ is a tunable privacy-loss parameter: for smaller values, it implies that the probability distributions of this mechanism acting on neighboring datasets are virtually indistinguishable (meaning more privacy) while for larger values it implies that the probability distributions are easy to discriminate between (meaning less privacy).

While there are many variants of differential privacy and many more implementations of these definitions, we rely on two core principles in differential privacy. The first is the centralized model, in which there is a data curator that publishes aggregate statistics on sensitive individual-level data. This model most closely resembles the paradigm under which the Census operates, so it is the one we consider here. Second, under the centralized paradigm one of the most fundamental mechanisms is the Laplace mechanism. This mechanism simply computes a statistic of interest, such as a mean, and adds random noise drawn from the Laplace distribution. This mechanism is guaranteed to be differentially private so long as the scale of the random noise drawn from the Laplace distribution is set to the global sensitivity of the mechanism divided by ϵ, where global sensitivity is defined as $\mid M(x) - M(x') \mid$.

2.2 Census' plan to implement differential privacy

While the Census has announced that it will implement differential privacy for its data products starting in 2020, the exact details of their implementations are still being developed. Furthermore, the Census' proposed implementation will likely introduce new developments in differential privacy, including how to deal with summary statistics for which there can be no added privacy noise (frequently called invariants) due to the organization's Constitutional mandate and imposing structural zeroes for counts of certain populations, like three-year old grandmothers [1].

The ACS poses even more challenges, including working with high dimensional data, generating geographic-specific estimates, preserving household associations, handling outliers in economic variables, and dealing with the ACS' survey weights [1]. At a fundamental level, differential privacy is designed to protect individuals, which has led some to raise concerns regarding the feasibility of generating credible small-area estimates [1]. With these challenges in mind, we develop a stylized example to produce conservative estimates of the expected effects of differential privacy on survey weighting. This is described in more detail in later sections.

2.3 Survey weighting

Since it is difficult for survey researchers to conduct simple random sampling on a large population of interest, they are often forced to obtain samples through other methods and are frequently left with samples that do not reflect the demographics of the population. In order to produce estimates for the population, it is thus desirable for survey researchers to compute weights that give more importance to responses from populations underrepresented in the sample and less importance to responses from populations overrepresented in the sample.

Many methods exist to compute survey weights. For a non-exhaustive list, see this summary compiled by the Pew Research Center [2]. The method that we focus on in this paper is cell weighting, which is a form of post-stratification weighting. For cell weighting, a researcher needs to have the full joint distribution of the population of interest. Using the joint distribution, they are able to compute weights that rebalance the sample to look more like the population with respect to the variables contained in the joint distribution.

Since it requires the full joint distribution of the population, cell weighting is not always feasible in practice. This is a non-issue when the population is U.S. adults since we can estimate the joint distribution using Census data products like the ACS. Furthermore, we think that differential privacy could affect methods like survey weighting by altering small cell counts in the joint distribution. This makes cell weighting a ripe candidate for this project.

3. Methodology

3.1 Data Aggregation and Processing

Every 5 years, the ACS publically releases Public Use Microdata Sample (PUMS) files containing a subset of individual respondents' records on various geographic levels.[1] Each record contains the age, marital status, gender, education level, and many other demographic features for the respondent. Additionally since each PUMS file is a subsample of respondents in a geographic region, a weight for each respondent is included to

[1] https://www.census.gov/programs-surveys/acs/technical-documentation/pums.html

indicate the number of people in the population that this given respondent represents. These records can be obtained on a granular block-group level or on a coarser state-by-state level. Using a simple R script (*pull_save_acs_pums.R*), we downloaded the 5-year PUMS files from 2012 for every state, keeping the age, citizenship status, marital status, educational attainment, sex, race, total income, and person weight. This data represents our estimate of the full U.S. population.

Our survey data comes from the Cooperative Congressional Election Study (CCES) conducted in 2016 (we are using the July 2017 release of this data).[2] The CCES is a large-N national survey conducted in every election year. It samples 50,000+ respondents in a nationally stratified manner, collecting the same demographic information as the ACS and respondents' answers to about 60 "common content" questions.[3] These questions survey respondents' behaviors, such as "In the past 24 hours, have you read a newspaper?" or "Do you use Facebook, Twitter or Instagram?", and their attitudes regarding controversial topics such as gun control, abortion, and taxes. To extrapolate the findings from the CCES to estimates for the national population, the survey employs a fairly complex sampling and weighting procedure based on population estimates aggregated from the ACS, the Current Population Study (CPS), and the Pew Religious Landscape Survey. To simplify our task, we decided to only generate survey weights for the CCES respondents based on their demographic data compared to the population estimates provided by the ACS.

Before we generated the cell counts for specific combinations of demographics in the ACS, we limited the demographic characteristics to state, race, education, sex, and age. Additionally, we binned the values/categories for race, education, and age in a consistent manner for the ACS and CCES datasets. This method of binning is standard practice in the survey-analysis community, so we chose to take advantage of it here to reduce the dimensionality of the cell counts in the ACS dataset.

Our methodology for binning race, education, and age follows. For race, every individual in the ACS and CCES are labeled as "White," "Black," "Asian," "Hispanic," or "Other." For education, individuals were labeled as "No HS" (did not graduate from high school), "HS graduate" (graduated from high school), "Some College" (attended some college or university), "2-year degree" (attained an associate's degree), "4-year degree" (attained a bachelor's degree), and "Postgraduate degree" (attained a graduate degree such as a Master's or Ph.D.). Lastly for age, individuals were labeled based on the age ranges less than 35 years-old, between 35 and 50, between 50 and 65, and older than 65 years-old.

After our processing steps, we obtained an ACS dataset with the state, race, education, sex, age and person weight for each individual, in which race, education and age had been categorized as described above. Our

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/GDF6Z0
[3] https://cces.gov.harvard.edu/

CCES dataset was processed in a very similar manner.

3.2 Cell-weighting

Once the ACS and CCES datasets were processed, cell weights for the CCES respondents were calculated. First, we computed five-way cell counts on the ACS data for each combination of state, race, education, sex and age. This produced 12,084 unique combinations among this set of five features. For each group, the person weights (the number of people in the population that this ACS respondent represents) of the individuals in that group were summed to estimate the total number people in the national population that have this specific combination of demographics. Second, we used the R "survey" package to calculate the weights for each respondent in the CCES based on the demographic cell counts from the ACS data. We use these as the "true weights" for the CCES respondents.

We also generated a second set of survey weights based on DP releases of the ACS cell counts, which we call "noisy weights" for the CCES respondents. To generate DP ACS cell counts, we added Laplace noise to each cell count. The Laplace distribution from which the random noise was chosen was centered at 0 and had a scale of $\left(\frac{2 \cdot 549}{\epsilon}\right)$. The global sensitivity is equal to $2 \cdot 549$ and ϵ is a specified privacy-loss parameter. For our mechanism, the global sensitivity when we used the person weights to compute the ACS cell counts is $2\cdot$ max(all person weights from all states) because each cell count is essentially an independent bin in a histogram. In the worst case, moving an individual with the most rare combination of demographics (person weight = 549) from one state to another would subtract that person's weight from the original cell count/histogram bin and add it to the new cell count/histogram bin. Thus, the difference between the two datasets would be $2 \cdot 549$. The obvious concern is that the person's weight would actually change when we move them to a new bin (because of the nature of what the weight is), but we are operating under the assumption that this weight would stay constant. Any DP ACS cell count that fell below 0 was given a count of 1 because we can not have negative cell counts. Once we had the DP ACS cell counts, we again used the R "survey" package to calculate "noisy weights" for the CCES respondents.

3.3 Testing effect of weights on CCES Survey Responses

Using both the "true cell weights" and the "noisy cell weights", we calculated the proportions of respondents that supported Trump or Clinton in the 2016 presidential election and the proportion of respondents that supported a national assault rifle ban. We used the R "survey" package to calculate the proportions for both topics, since it also allows us to calculate these proportions based on specific demographics like race and education.

4. Results

We perform several simulations to evaluate the impact of releasing a differentially private version of the 2012 5-year ACS estimates - the same data to which the 2016 CCES was actually weighted - on the computation of post-stratification weights for the 2016 CCES and the subsequent estimates for various survey items. We focus on two survey items: vote preference for the 2016 general election and support for an assault rifle ban. For the vote preference item, our estimand is the difference between the proportion of respondents who say they will vote for Hillary Clinton and the proportion of respondents who say they will vote for Donald Trump (for brevity, this will be called the Clinton lead). For the assault rifle ban item, our estimand is the proportion who support the ban minus the proportion who oppose the ban, or the net support for an assault rifle ban.

Our primary concern is that differentially private releases of population data could introduce relatively large amounts of error into the ACS cell counts for smaller populations and cause the resulting post-stratification weights to be incorrectly calibrated for some demographic subgroups. If this is the case, we would also expect that CCES estimates using those weights would vary from the ground truth estimates when the weights are computed to non-noisy ACS data. With this in mind, we choose to report the results of subgroup estimates for these two survey items. We compute estimates for Clinton's lead for each race category and estimates for the net support for an assault rifle ban for each education category.

Since we are using the Laplace mechanism to add noise to the ACS cell counts, we compare the weighted estimate obtained from the differentially private release of the ACS data to the weighted estimate obtained from the true ACS data for different values of the privacy loss parameter, epsilon. The results of our simulations are summarized in Figure 1 and Figure 2.

For the largest demographic subgroups (e.g., whites and people with no high school education), the effect of differential privacy on these estimates is negligible. Even with an ϵ of 0.1, the noise added to the ACS cell counts does not affect the post-stratification weights drastically enough to change Clinton's estimated lead among whites or net support for an assault rifle ban among those without a high school education. This is the behavior we expected here since applying differential privacy is not expected to dramatically affect summary statistics when there is a large sample size.

For smaller subgroups, we do observe some shifts in the CCES estimates for smaller values of epsilon. For

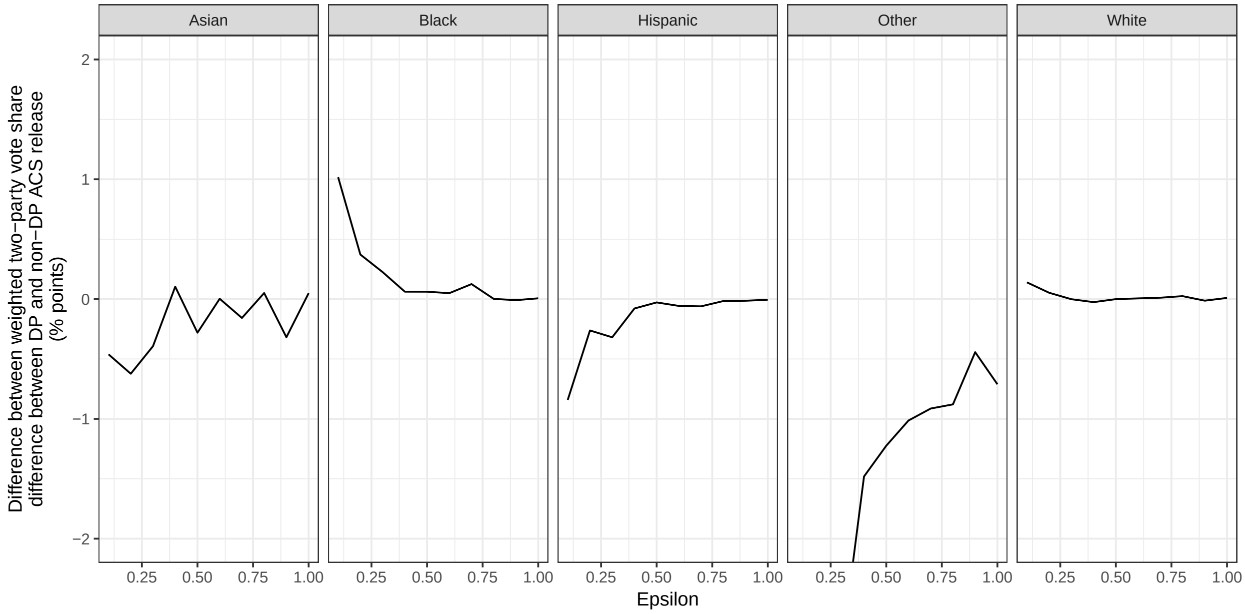

Figure 1: Evaluation of estimates of Clinton's lead in the CCES using weights computed from hypothetical private and non-private ACS releases for various values of epsilon.

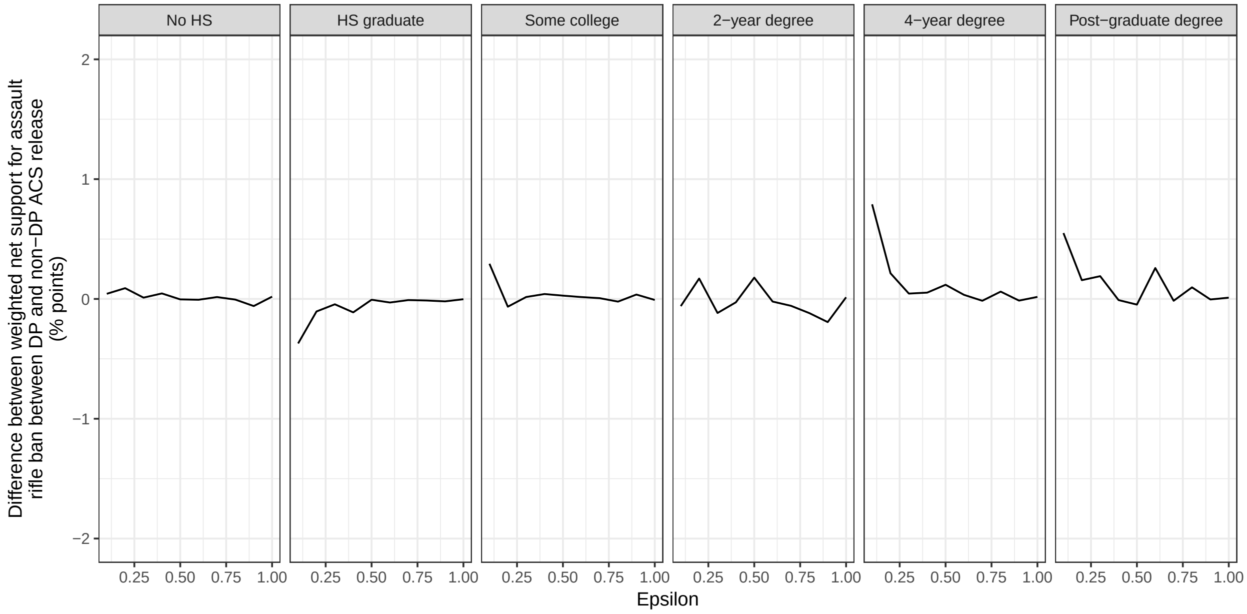

Figure 2: Evaluation of net support for assault rifle ban by education in the CCES using weights computed from hypothetical private and non-private ACS releases for various values of epsilon.

instance, using an ϵ of 0.1 to add Laplace noise to the ACS cell counts and then weighting the CCES to this data moves the estimate of Clinton's lead among black Americans by roughly a percentage point. Similarly, with an epsilon value of 0.1, the estimate of net support for an assault rifle ban among those with a 4-year degree (who account for about one-third of the U.S. population) shifts by about one percentage point when the post-stratification weights come from the noisy ACS data. As we increase the value of epsilon and decrease the amount of noise we add to any cell count, the difference between the estimate obtained from weighting the survey to the private and non-private population data goes to zero.

5. Conclusion

Our analysis demonstrates that even with the addition of maximum Laplace noise, the population-level estimates of presidential candidate preference and support for gun control do not change much for various large demographic subgroups. However, for small values of ϵ the estimates move by as much as one percentage point for some subgroups. On its own, this may not be particularly meaningful, especially since survey researchers are accustomed to dealing with other forms of error, including sampling and measurement error. But this privacy error, even if it is small, is still a new type of error that researchers will need to incorporate into their estimates.

Additionally, this analysis is still a relatively conservative one. The concern we have expressed is that adding noise to population data would affect the relative count of small subgroup populations. But isolating just race or just education does not leave us with particularly small sub-populations, which means that we might observe more substantial and meaningful shifts for combined education and race groups, for example. Our analysis, however, is a stylized one due to uncertainty that exists in how the ACS will implement differential privacy and to our desire to simplify the CCES sampling and weighting schemes for illustrative purposes in this paper. We choose not to evaluate these smaller subgroups out of the fear that any effects we discover would be artifacts of our stylized simulation and not of the expected effects of differential privacy on statistical calibration techniques. Thus, our analysis is conservative but it is so by design.

Despite the observation that the noise we are adding does not affect our data, it would be interesting to implement and test how clipping the ACS "person weights," especially the larger weights, would affect the ACS cell counts and corresponding cell weights. Clipping is a mechanism that is generally implemented in a differentially private mechanism to decrease the global sensitivity and protect outliers in the dataset. It allows for less noise to be added to the DP-released statistic, which should increase the utility of the statistic. How clipping to different values for "person weight" biases the ACS counts and cell weights for certain demographics would need to be studied to determine the optimal clipping value. Furthermore, exploring the effect of clipping on specific questions asked in the CCES would be useful to examine the introduction of any biases into the post-stratification weighting.

The PUMS data released by the ACS is subset of the population for a certain geographic level (in our case on the state level). Ideally, ACS would use a DP synthetic data generation process to release the PUMS data. We would have liked to simulate this synthetic data generation process, but this would have required having access to the full PUMS data, which is not possible. Adding Laplace noise to the demographic cell counts generated from the PUMS data is a workaround to study the effects of a DP synthetic data release on surveys that rely on the ACS to adjust the findings in a sample. This stylized study provides a springboard from which to further explore how the a DP ACS data product will affect downstream statistical calibration methods.

6. References

[1] Dajani, Lauger, Singer, Kifer, Reiter, Machanavajjhala, Garfinkel, Dahl, Graham, Karwa, Kim, Leclerc, Schmutte, Sexton, Vilhuber, Abowd. 2017. The modernization of statistical disclosure limitation at the U.S. Census Bureau.

[2] Mercer, Lau, and Kennedy. 2018. " For Weighting Online Opt-In Samples, What Matters Most?" Pew Research Center. Accessed from: https://pewrsr.ch/2HdVXfb.

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

[4] Wood, Alexandra, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, et al. 2018. Differential Privacy: A Primer for a Non-Technical Audience. Vanderbilt Journal of Entertainment & Technology Law 21 (1): 209.