The Problems of Minimal Support: Considerations for an Establishment Survey of Local Election Officials

Paper Presented at the Annual Meeting of the
Southern Political Science Association, San Juan, PR

Jay Lee[*]
Paul Gronke[†]

2020-01-05

Abstract

In this paper, we provide evidence to support the use of a specific sampling algorithm for drawing random samples of local election officials (LEOs) in the United States, using the `sampling` package in the `R` statistical package. The paper is part of a larger project that examines the backgrounds, professional orientations, and opinions of LEOs in order to better characterize their role as "stewards of democracy." The enormous diversity of local jurisdictions and the hyperfederalized institutional structure of American elections combine to create methodological challenges to drawing a random sample that allows generalizations both about LEOs and also about the American voting experience. The paper explores the statistical foundations of a number of unequal inclusion probability sampling methods implemented in the `sampling` package. We show using simulations that the extremely skewed distribution of jurisdictions (by population size) causes anomalies in the sampling method, resulting in overly variant samples and extreme values for sampling weight when using the minimal support sampling algorithn. We further show that the "random systematic" sampling method is superior, resulting in lower variance estimates, and is just as easy to implement as "minimal support".

Acknowledgements: The authors would like to thank our collaborators on the DF/RC LEO Survey, Natalie Adona (Assistant Clerk-Recorder/Registrar of Voters, Nevada County, CA), Sarah Cole (Research Assistant, Democracy Fund), and Professor Evan Crawford (University of San Diego).

The 2018 and 2019 DF/RC LEO Surveys were supported by the Elections Program of the Democracy Fund, Adam Ambrogi, Program Director. Reed College also provides support to the Early Voting Information Center. Tammy Patrick, Senior Advisor to the Elections Program at the Democracy Fund, and a group of state and local election officials around the country, provided invaluable feedback on our questionnaire and encouraged local officials to respond. The DF/RC Advisory Board helped us work through sampling problems; Professor David Kimball (University of Missouri, St. Louis) was especially responsive.

This research rests on the foundation of the LEO survey research conducted by Dr. Eric Fischer of the Congressional Research Service and his research collaborators. Welcome to retirement, Eric! Rebecca Gambler, Director, Homeland Security and Justice at the Government Accountability Office, provided helpful insights into the data collection procedures used by the GAO in their 2018 report. David Kuennen, Senior Research Program Specialists at the Election Assistance Commission, the Elections Research Team at Fors/Marsh, and Susan Dzieduszycka-Suinat, President and CEO of the U.S. Vote Foundation, helped us navigate between different lists of elections jurisdictions.

Paul Manson, EVIC Research Director, has been a constant source of support, guidance, constructive criticism, and bags of popcorn. We owe him a hearty thank you and a few tropical beverages.

All interpretations and conclusions are the responsibility of the authors, and do not represent the views of the Democracy Fund or Reed College.

[*]Early Voting Information Center, Reed College, Portland, OR.
[†]Early Voting Information Center and Professor of Political Science, Reed College, Portland, OR

Introduction

Local election officials (LEOs) in the United States have been described as the "stewards of democracy" (Adona et al. 2019), the street-level administrators who play a central role in connecting American citizens to elections and to their democracy. In 2018 and 2019, the Early Voting Information Center at Reed College, in partnership with the Elections Program at the Democracy Fund, conducted two nationwide surveys of local election officials to ask them about election administration, access, integrity, and reform ("Survey of Local Election Officials" 2019). The broad goal of the DF/RC LEO Survey was to elevate and amplify the voices of local election officials in discussions and debates over election integrity and election reform. These "street level administrators" of American elections play a critical role in assuring a free, fair, and secure elections system, and their viewpoints and opinions about election administration deserve to be heard.

We also entered this project with scholarly interests in the LEO as a bureaucratic and political position in local government. There are local election officials in every United States county, in every parish in Lousiana, and in every township and municipality in Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, Rhode Island, Vermont, and Wisconsin. We want to learn about who takes on the role of an LEO, what their professional environment is like, and how they embrace their role in protecting the integrity of American elections and improving trust and confidence in our democratic system. Finally, we care about way to improve the voter experience in the United States. Consequently, we pay attention to how laws, administrative procedures, and decisions made by LEOs influence voter information and attitudes about election integrity, the accuracy of the vote count, and the resilience of our system in the face of domestic and foreign security threats.

Simultaneously considering the **local official** and the **local voter** creates two important decision points that impact sampling and substantive interpretation of the results. This paper focuses on the mathematical and statistical sampling implications of these decisions and provides recommendations for future research.

The first decision point is linked to federalism and to definitions about the meaning of "election jurisdiction" and "local election official." In pratical terms, this boils down to a decision about how to treat Minnesota. Some research projects collect Minnesota information at the county level while others collect information at the municipal and township level. This decision is an easy one to make, but still consequential, because it changes the size of the sample universe by 25%. We briefly discuss the meaning of the term "local election official" and how it provides guidance to the choice of how to treat Minnesota.

The second decision point is more difficult to adjudicate. While it is also a product of federalism, is more far-reaching than deciding how to handle a single state. The highly diverse and decentralized American elections system has been described as "hyperfederalized" (Ewald 2009) and a "complex quilt" (Gronke 2014). The hyperfederalized nature of American elections makes it difficult to generalize about a single American election ecosystem, draw a profile of a typical local election official, or make assertions about a common voter experience. Some aspects of election administration are governed by the Constitution and by federal statutes, but the administrative details for federal, state, and local elections are largely determined by fifty separate state governments, even though local officials may have substantial administrative autonomy.[1] The responsibilities of LEOs can range from determining political boundaries, establishing eligibility for exercising

[1] Local jurisidictions are creatures of and subordinate to state governments, except in "home rule" states, and it is not clear how much home rule includes election administration. The legal doctrine in play is "state sovereignty" and "Dillon's rule", established by the *Hunter v. Pittsburgh* decision. See Burns (1994) for an extended discussion of the legal and political status of local governments.

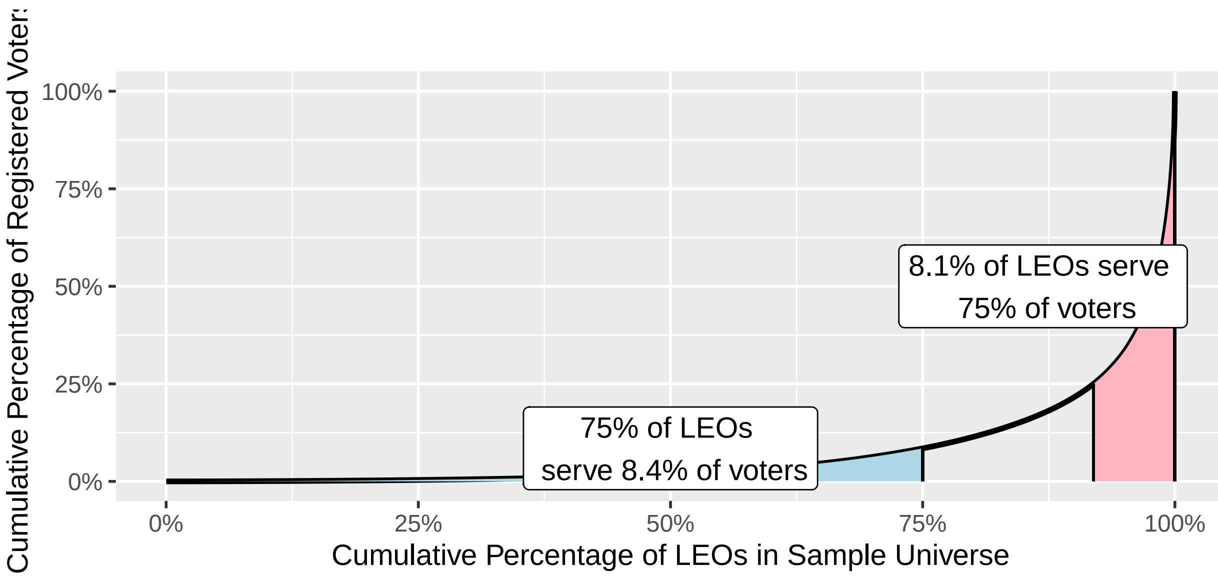

Figure 1: Illustrating the 75:8 Problem in United States Election Administration

the franchise, selecting appropriate election technology and election infrastructure, conducting early and election day voting and processing mail ballots, certifying election results, and conducting post-election audits. Yet, because state election codes are so important, we have to pay close attention to how our sample represents different states.

A related contextual feature that has to be considered is what is referred to in the field as the *size principle* or the *80:20 problem*, though a more accurate term may be the "75:8 problem". Local election official in the United States vary enormously in the numbers of voters that they serve. As shown in Figure 1, 75% of LEOs serve 8.4% of voters, while 8.4% of LEOs, serving in large jurisdictions, serve 75% of voters. In addition, a small number of states contain a large number of the smallest jurisdictions and a small number of jurisdictions (in a small number of states) serve large numbers of voters. The consequence is that a randomly chosen set of "small", "medium", and "large" jurisdictions may not be representative of the population in those categories. While we may not make assertions about a single American voter experience, we want a good sense of the **range** of voter experiences and how the voting experience can be improved by the actions of LEOs.

The implication is that while our research is focused on local election officials, sampling occurs among local election jurisdictions. These local jurisdictions are distributed unequally across the fifty states and the District of Columbia. All of these different levels are substantively important. Our sampling procedure has to assure a sufficient distribution across states, among LEOs, and among service populations (i.e. registered voters).

A typical solution in situations like this, where an important quality of the population is highly variant or highly skewed, is to use unequal probability sampling, choosing sampling probabilities relative to the size of the underlying unit (i.e. larger units are sampled with much higher probability). However, there is more than one method of sampling using unequal inclusion probabilites, and the mathematical and empirical question we face is: which to use?

We are still in the process of working out the technical portions of our work, but in the third section of the paper, we present a first cut at theoretical results, based on mathematical derivations and simulations. Within the field of unequal probability sampling, there are multiple mathematical techniques (algorithms) that can draw a sample proportional to size. In 2018, we used the the "minimal support sampling" method implemented in the `sampling` package in `R` (Tillé 2016, 2010) because it requires no prior assumptions and was straightforward to implement. However, the problem we discovered in the second wave of our survey was that the minimal support method was not replicable, i.e. we could not reproduce our 2018 sample.[2] Simulations revealed that in the presence of a very skewed distribution, the minimal support method creates potential anomalies by over-sampling one or the other ends of the population distribution. We have not been fully able to mathematically demonstrate this phenomenon, but the results of our simulations are consistent. Another sampling method—the random systematic method—avoids these problems.

There is a second problem that we encountered that is common to all unequal sampling methods: very large survey weights.[3] These are a particular issue when surveying LEOs because of the federalist overlay. A small jurisdiction in Michigan or Texas, for example, simply can't "stand in" or "represent" a small jurisdiction in Connecticut or Oregon. While excessively high survey weights are a point of concern that we are still investigating, the minimal support method, because it occassionally oversamples small cases, sometimes ends up with too many sampled cases with large weights and, consequently, highly variant weighted means. Even though, in the long run, weighted means from samples based on the minimal support method converge on the population mean, the chance for an anomalous sample is far too high for us to recommend this method. Our simulations show that on data that share the specific characteristics of the LEO population, the minimal support method produces these kind of samples far more frequently than the random systematic method.

There is no perfect way to sample the population of LEOs. The federalist overlay, and the extreme skew in the service population, make it a complex enterprise. Ultimately, we hope in this paper to evaluate the downsides of unequal probability methods (large weights) along with any critiques of other methods (which we have not yet included in this work). We are able to conclude at this stage is that one specific type of unequal sampling (minimal support) is worse than other types, and we demonstrate this in the final section of the paper.

Identifying the Universe of Election Jurisdictions and Local Election Officials

The starting point for any research project focused on local election administrators is to identify and create a list of the population of LEOs, and subsequently develop a sampling design that produces a statistically valid sample with responses that can be generalized to the population of LEOs and their service populations. We adopt in our research a definition of "local election official" proposed by Hale, Montjoy, and Brown (2015), p. 2-3: "local election officials are the local officials with statutory responsibility for elections at the local level." While this definition is straightforward, American federalism and diversity combine to create a series of decision points about what constitutes the universe of LEOs in the United States.

[2] Or rather, a sample drawn with a new random seed was notably dissimilar to our 2018 sample.

[3] For the uninitiated, sampling weights correct for imperfections in the sample, such as discrepancies in coverage or non-responses, that could lead to discrepancies between the sample and the reference population. They are essentially multipliers that allow one randomly selected observation to "stand in" for others.

To start, there is no single, comprehensive list of local election jurisdictions, or the civil servant (or servants) who has (or have) statutory responsibility for conducting elections in that jurisdiction, and the existing literature does not use the same list, and uses different procedures for sampling within that list.[4] Depending on how you count, there are between 7,858 (Kimball and Baybeck 2013) and 10,370 (Kimball et al. 2010; Government Accountability Office 2018) local jurisdictions with the primary responsibility for conducting elections in the United States. As already mentioned, the different totals occur because in some states, there are sub-jurisdictions that have responsibility for some aspects of election administration. Notably, one state—Minnesota—lodges responsibility for polling place administration (e.g. deploying election technology, hiring and training poll workers) at the sub-jurisdictional unit. Studies that use the larger N consider these sub-jurisdictional units as the basic unit of analysis. The fact that the total varies is an early indicator of the inherent risks in trying to make sweeping conclusions about election administration and election administrators in the United States. American elections are administered by the local election officials in counties, municipalities, and townships throughout the county. Hale, Montjoy, and Mitchell (2015) write that: "…the actual conduct of elections is primarily a function of local governments," while Alex Ewald, in his historical study of American election administration, reminds us that "the United States has always placed responsibility for running national and state elections in the hands of city, town, and county officials" (Ewald 2009). The total number of jurisdictions also varies depending on what "statutory responsibility" you are interested in examining. We have encountered the higher total when the interest is in deployment of election technology on Election Day (and in some cases, for pre-election day voting). For example, the Government Accountability Office's 2018 report, "Voting Equipment Use and Replacement," reports 10,340 election jurisdictions in the United States (Government Accountability Office 2018). The GAO conducted a survey of the "chief election official within the jurisdictions selected" to understand the choices made about voting technology and polling place management.[5]

In summary, when trying to generalize about election administration in the United States, there are (at least) three important issues to consider:

1. Legislating and rule making are done primarily at the state level, so a representative sample will need to assure a sufficient distribution of responses across states.
2. The actual administration of elections is done at the local level, and the granularity of "local" varies across states. In 42 states, elections are administered at the county level, but in 8 states, elections are administered at the village, township, and municipality level. This has an important impact on how we sample. Among the 7,800 LEOs in the US, 3,400 work in just two states — Wisconsin and Michigan — and another 1,500 administer elections in towns and villages in New England. A simple random sampling scheme will therefore vastly over represent these states, even though it may do a good job representing the population of all LEOs.
3. Finally, voters are not evenly distributed across jurisdictions; Figure 2 shows the significant right skew in registered voters by jurisdiction. One jurisdiction alone — Los Angeles County — contains 3.2% of

[4]We use the plural because in Missouri, there are two individuals, one Democratic and one Republican, who are appointed to act at the local election officials for each county. A number of other states have local election boards, but there is usually an individual, sometimes the head of the board, other times a separately appointed individual, who can be identified as the administrative lead.

[5]See Appendix I of the GAO report for a detailed description of their sampling strategy, also quoted below. We specifically do not use the term "chief election official" because that term refers to official in charge of elections at the state level (Hale, Montjoy, and Brown 2015, 3), a usage we have confirmed in personal communications with a number of field matter experts. Kimball Brace, president of Election Data Services, similarly reports 10,071 as the total number of LEOs in 2013 (cited in Kropf and Pope 2019, 188), possibly because Brace tracks and maps the use of election technology across the United States.

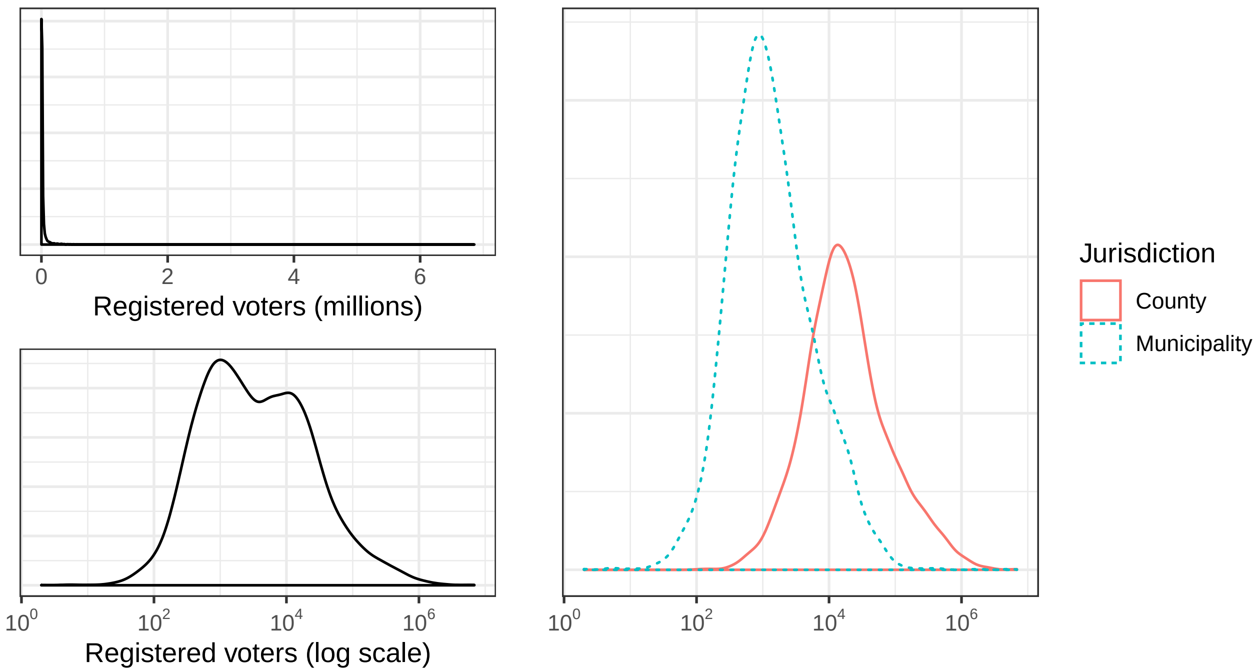

Figure 2: Distribution of registered voters by jurisdiction

ALL registered voters in the country. The largest 10% of jurisdictions serve almost 80% of registered voters. If we are interested in not just generalizing to LEOs, but to LEOs who administer elections to a substantial portion of the American public, we must take voter totals into account.

Our final sample universe of 7,834 LEOs is 2,000 less than the GAO total, and much closer to the total of 7,858 reported by (Kimball and Baybeck 2013), with the difference almost completely a function of Minnesota.[6] We consolidate some additional jurisdictions in our sample because of choices that were made in New Hampshire, which reports data to the EAVS in some cases at units below the MCD level. For example, Concord, NH results are reported in 10 different lines in the EAVS (each ward separate), while we consider Concord, NH to be a single administrative unit with a single LEO. Our target sample size was 3,000, assuming a response rate of approximately 30% (the response rate achieved by Kimball and Baybeck). Readers should be aware of the variations in the sample universe when comparing across studies.

Limitations of the Sampling Methods Used In Previous Surveys

Once our population universe is defined, we need to decide upon a method for choosing our sample. Population variation within the sample universe has a significant mathematical impact on any results that we might find—a simple random sample will include lots of small jurisdictions and very few large jurisdictions. The size of the large jurisdictions makes them impactful to more voters, however, so it might be sensible to sample all of the largest jurisdictions to ensure that we have results representing this large chunk of voters.

[6]Like Kimball and Baybeck, we sample at the county level in Minnesota and at the MCD (minor civil division) level in Wisconsin and Michigan, relying on past research that shows that most administrative decisions are made by township or municipal clerks in the latter two states, and by county administrators in Minnesota. See (Huefner et al. 2007), cited in Kimball and Baybeck (2013), p. 131.

Surveys dealing with this sort of skewed population include establishment surveys (e.g. surveys of hospitals, universities, governments, etc.), which are common in economics research (DesRoches 2011), as well as surveys dealing with land use or geographical area ("Sampling Techniques" n.d.).

Previous surveys ("How Local Election Officials View Election Reform: Results of Three National Surveys" 2011; Kimball and Baybeck 2013; Kimball et al. 2010) recognize these facts, and most have adapted stratified sampling methods with manually-defined size bins. For example, a seminal set of surveys were administered after HAVA, overseen by Eric Fischer of the Congressional Research Service and Professors Kevin Coleman and Carol Silva, stratified by "size," using the number of jurisdictions as a surrogate for size (See ("How Local Election Officials View Election Reform: Results of Three National Surveys" 2011), Fischer and Coleman (2005), and Moynihan and Silva (2008)):

> The three surveys were conducted after the November 2004, 2006, and 2008 federal elections, between December and the following April. For each survey, a sample of approximately 3,800 LEOs was drawn from the roughly 9,000 election jurisdictions in the 50 states. To ensure that LEOs from all states were included, but that states with large numbers of LEOs were not disproportionately represented (see Figure A-1), a modified random-sampling regime was used, as follows: Surveys were sent to all LEOs in states with 150 or fewer local jurisdictions. For the ten states with more than 150 LEOs, a sample of 150 was chosen at random from the local jurisdictions, and surveys were sent to those LEOs. ("How Local Election Officials View Election Reform: Results of Three National Surveys" 2011, pg. 81)

This sampling procedure appropriately adjusts for the differences across jurisdiction size, and the overwhelming number of small jurisdictions, but using an awkward algorithm that is based in part on historically contingent decisions about drawing county lines.

Kimball and Baybeck (2013) advanced upon the CRS and served as a model for our research. They developed a sampling rule that was not connected to specific states (or the number of jurisdictions within states), but is tied directly to jurisdiction size:

- 100% inclusion probability for jurisdictions with a registered voter population above 50,000;
- 50% inclusion probability for jurisdictions with a registered voter population between 5,000 and 50,0000;
- 25% inclusion probability for jurisdictions with a registered voter population below 5,000.

It is illustrative to read their rationale for these decisions at length:

> To simplify some of the analyses that follow, we divide the universe of local jurisdictions into three size categories: small (serving less than 1,000 voters), medium (serving between 1,000 and 50,000 voters), and large jurisdictions (serving more than 50,000 voters). We chose 1,000 voters as one dividing line because jurisdictions with fewer than 1,000 voters are generally small towns that have no more than a couple of polling places and a handful of poll workers. We expect these jurisdictions to have a different election administration experience than larger jurisdictions. In addition, roughly one-third of the jurisdictions served less than 1,000 voters in recent presidential elections, so this serves as a natural break in the data.
>
> We chose 50,000 voters as the other dividing line because jurisdictions serving more than 50,000 voters tend to be in densely populated metropolitan areas with a large central city. Thus, the largest jurisdictions have different infrastructure and transportation networks than the medium-sized jurisdictions, which are mostly rural and exurban counties. Together, these dimensions characterize what we define as small, medium, and large jurisdictions in a variety of analyses below. The smallest jurisdictions are primarily in the upper Midwest and New England, with a smaller number in the Plains. Large jurisdictions are concentrated in the major metropolitan

centers of the United States. (Kimball and Baybeck 2013, 132)

Within those three strata, they sampled all 462 units in the top stratum, 2000 from the 4619 medium sized, and 500 from the smallest. They made these choices in order to "to ensure representation of the largest jurisdictions" (2013 p. 132) and to have the ability to make generalizable statements about LEOs serving in medium- and small-sized jurisdictions. This sampling algorithm improves upon the CRS, but because it relies on large population categories, it still has the undesirable feature of sampling a unit with a registered voter population of 49,999 with the same probability as a unit with 1/50th the size.

A third and more sophisticated version of the Kimball and Baybeck method was developed by the Government Accountability Office (Government Accountability Office 2018).

> To obtain a representative sample that included a mix of both rural and non-rural jurisdictions, we used a two-level stratified sampling method in which the sample units, or jurisdictions, were broken out into rural and non-rural strata. ... Of the 10,340 local election jurisdictions nationwide, 70 percent were classified as non-rural while 30 percent were classified as rural. ... We selected a two-level stratified sample of 800 local election jurisdictions. Using the RUCC [rural and urban] codes, we allocated 600 sampling units, or jurisdictions, to the non-rural stratum and 200 to the rural stratum. To obtain a sample that also reflected the population distribution across jurisdictions nationwide, we used the population of the local election jurisdiction as the measure of unit size and selected the sample units within each stratum with probability proportionate to population of the local election jurisdiction, without replacement. ... Because the sample was selected with probability proportionate to population size, any jurisdiction (county or MCD) with more than about 225,000 people was selected with certainty.

Like Kimball and Baybeck, the GAO recognizes the different administrative environment faced by administrators in rural and urban jurisdictions. In order to do this, they stratified their sample by "rural" and "urban" counties and "minor civil jurisdictions" (the MCD–level states are Connecticut, Maine, Massachusetts, Michigan, Minnesota, New Hampshire, Rhode Island, Vermont, and Wisconsin), and then sampled proportional to population within these two strata. When applied to their sampling universe, this meant that at some cutpoint, units were sampled with a probability of 1.0.

The Kimball/Baybeck and GAO sampling procedures resulted in a desirable set of jurisdictions in the final sample (although there are some details in the CRS samples that are difficult to unpack), but there are problems that we would like to avoid. In the case of the CRS, relying on jurisdiction count rather than jurisdiction size causes some disparate units to have the same inclusion probability, such as Harris County, TX, with a registered voter populations of 2,258,976, and King County, TX with a registered voter population of 176. Kimball and Baybeck perform much better, but their algorithm still treats the endpoints of their categories identically—a jurisdiction with a registered voter population of 49,999 is sampled with the same probability as one with a population 1/50th as large.

We wanted to develop a mathematically rigorous method of sampling proportional to the population of registered voters which does not rely on a manually-defined formula for stratifying (it may be that the GAO technique of stratifying by rural/urban may work well, and avoid the mathematical anomalies we identify in the rest of this paper). An additional advantage of developing a mathematically rigorous method of sampling is that it will allow us to take better advantage of statistical techniques like hypothesis testing and confidence intervals, and should be replicable across studies. In other respects, our sampling design follows in spirit and in logic the sampling approach and presentation guidelines developed by Kimball and Baybeck (2013) and the research team at the GAO, and like the GAO, we rely on a method with unequal probabilities of sampling. Like those research teams, we are cautious about interpreting weighted means without paying close attention to variation across jurisdiction size.

Unequal probability sampling

In any sampling method, each unit of the population is assigned an *inclusion probability*—-the probability that an individual unit will be included in a sample. For simple random sampling (or any equal probability sampling method), this inclusion probability is the same for every unit k. An *unequal probability* sampling method is a sampling method in which the inclusion probabilities are not identical for every unit in the population. A stratified sampling method can be an unequal probability method, if the inclusion probabilities differ across strata.

Mathematically, the size of jurisdictions in the United States follows a log-normal distribution, i.e. the natural logarithm[7] of the observations is approximately normally distributed, which has a significant right skew (nearly 80% of American voters live in the largest 10% of jurisdictions).[8] One solution for a highly skewed distribution is to sample proportional to unit size. This is often referred to as PPS sampling: probability proportional to size (Lavrakas and Cohen 2011). Sampling among unequally sized entities is a known issue in the sampling literature, and is typically referred to as "establishment sampling" (examples include sampling firms or schools). The method we employ derives an unequal probability of selection based on the size of the service population, in this case, the number of registered voters.[9]

In the following sections, we explore the mathematical and statistical foundations of these methods, and show how sampling relative to unit size, using first order inclusion probabilities (the foundation of the "minimal support" method), may interact with the underlying sampling formula in such a way that the largest or the smallest units are badly under sampled.[10]

Use cases

Two typical use cases of unequal probability sampling are for establishment surveys and multistage sampling.

- An *establishment survey* is a survey where the unit of analysis is some sort of establishment: hospitals, universities, governments, etc. For example, a survey attempting to find out the total electricity consumption in a state might want to over sample the largest counties, as the impact of one high-consumption county on the total is greater than the impact of one low-consumption county.
- *Multistage sampling* is a sampling method that consists of multiple stages, often having a between-strata selection stage and a within-strata selection stage. If not every population stratum is to be represented in the sample, then a selection of strata is made using PPS sampling on the size of each strata. Within

[7]Or the logarithm with any base.

[8]See Figure 2 previously, and Figure 9 in the Appendix.

[9]The GAO used the population size, because "Census data were readily available for all counties and MCDs nationwide; we were able to find registered voter totals for nearly all counties and MCDs. Our thanks to the Director of Elections for the State of Michigan, whose office provided us registered voter totals for MCDs for the state.

[10]In our 2018 survey, we opted for the "minimal support method" that we are now advising against (Gronke, Paul and Crawford 2019). At the time, to confirm the validity of that sampling method, we compared the percentage sampled in each of the "Kimball-Baybeck" and "GAO" size categories, and the proportions were very close. Attempts to replicate the sampling formula prior to the 2019 survey revealed that we benefited literally from the luck of the draw. Simulations using the minimal support method, illustrated below, show that samples could have easily deviated substantially from an optimal mixture of small, medium, and large jurisdictions.

Probabilities	Replacement	Expression
Equal	With	$1-\left(\frac{N-1}{N}\right)^n$
Equal	Without	$\frac{n}{N}$
Unequal	With	$1-\left(1-\frac{x_k}{\sum_{i\in U} x_i}\right)^n$
Unequal	Without	Equation 6

Table 1: Common inclusion probabilities for standard classes of sampling methods

each of the sampled strata, a subset of cases are sampled for analysis. An example of this might be in selecting public school teachers to survey. The schools are selected using PPS based on the number of teachers (between-strata), then within each school we might conduct a SRS of individual teachers (within-strata).

Like any other method, there are both practical and mathematical reasons for using an unequal sampling method. When the variable of interest is correlated with size, PPS can provide greater efficiency (lower variance) than SRS. Even if this is not the case, it may be more practically convenient to employ unequal probability sampling. The school teacher example might be one such case - sampling a subset of schools (and oversampling large schools) will reduce survey costs compared to a genuine SRS of the population of teachers.

Inclusion probabilities

There are a few different ways to select appropriate inclusion probabilities for a PPS method without replacement,[11] but the literature has generally stuck to one specific formulation. We call this the "linear method", and detail its construction in the Appendix. There we additionally describe another formulation of inclusion probabilities (the "sequential method") and contrast the two methods. For many populations, the linear method will produce a *cutoff point*, at which every observation with a size larger than the cutoff point will be included in the sample with probability 1. As shown below, in our LEO population the cutoff point is 16,808. For comparison, Kimball and Baybeck (2013) had a cut-off point of 50,000 , and the GAO identified a cut-off point in their sampling – "Because the sample was selected with probability proportionate to population size, any jurisdiction (county or MCD) with more than about 225,000 people was selected with certainty" (Government Accountability Office 2018, 57).

Figure 3 shows the relationship between inclusion probabilities (from the widely-used linear method) and our size variable of registered voters.[12] The inclusion probabilities are directly proportional to size until the cutoff point is reached, at which point all of the inclusion probabilities are 1.

[11]Methods with replacement typically have higher variance in estimates, and are often avoided as such. Typical inclusion probabilities for other methods (with replacement, or equal inclusion probabilities) are included in Table 1.

[12]The x-axis of the graph was truncated at 50,000, to better illustrate the relationship in question. On a scale of 0 to 6 million, the diagonal line here appears vertical.

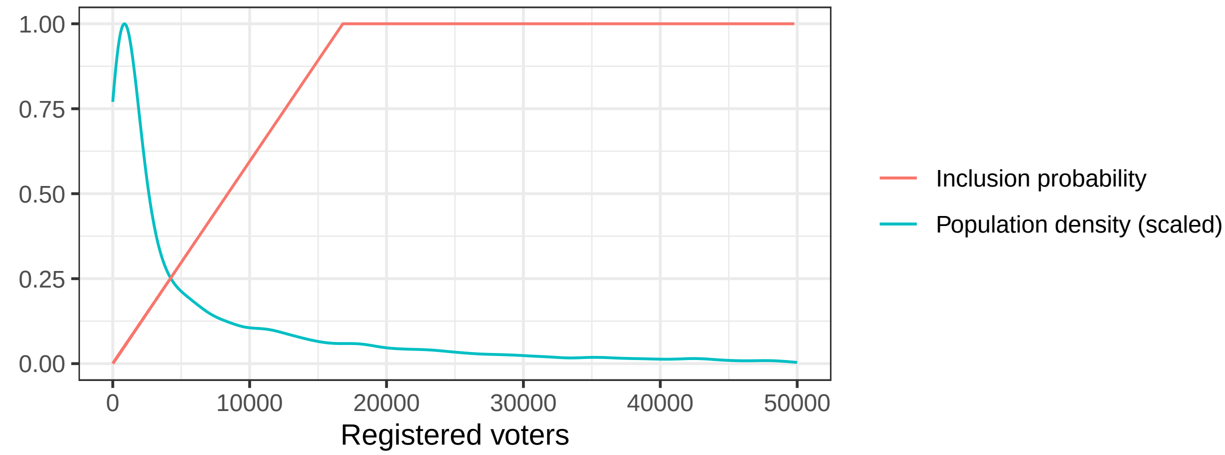

Figure 3: Density and inclusion probabilities of LEO population

Methods for analyzing an unequal-probability sample

As briefly mentioned earlier, the (unweighted) sample mean is a *biased* estimator of the population mean under an unequal sampling scheme; i.e. we do not expect a mean of sample means to approximate the true population mean. In fact, it should be biased towards the cases with larger inclusion probabilities.

One unbiased estimator for the population mean of a variable Y under an unequal sampling method (without replacement) is the Horvitz-Thompson estimator (Horvitz and Thompson 1952):

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{k/\in s} \frac{y_k}{\pi_k}. \tag{1}$$

This is a weighted mean, with a weight (as applied to the sample mean) on each unit of w_k as follows.[13]

$$w_k = \frac{n}{N\pi_k}. \tag{2}$$

This weight "re-balances" our sample (which over sampled large cases), by assigning a larger weight to the cases that were given a smaller inclusion probability.

While we can often account for a known bias through weighting, accounting for variance is harder. For fixed n, the variance of the Horvitz-Thompson estimator is

$$V\left(\hat{\mu}_{HT}\right) = \frac{1}{N^2} \sum_{i<j}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2, \tag{3}$$

and

[13] Some of these quantities presume that the population size N is known in analysis (though this can be estimated, see Overton and Stehman (1995)), or require that the sampling probabilities π_k are known for all units k of the population.

$$v\left(\hat{\mu}_{HT}\right) = \frac{1}{N^2} \sum_{i<j}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{4}$$

is an unbiased estimator for $V\left(\hat{\mu}_{HT}\right)$ (Sen 1953; Yates and Grundy 1953).[14]

The above listed variance will be small when

- The π_k are proportional to the y_k, and/or
- The values for $\pi_{ij} - \pi_i \pi_j$ are small (Sunter 1977).[15]

Assuming normality via the Central Limit Theorem,[16] we can use these to derive a 95% confidence interval for the true population mean of Y:

$$\left[\hat{\mu}_{HT} - 1.96\sqrt{v\left(\hat{\mu}_{HT}\right)},\ \ \hat{\mu}_{HT} + 1.96\sqrt{v\left(\hat{\mu}_{HT}\right)}\right] \tag{5}$$

PPS compared to other methods

We earlier mentioned that unequal probability sampling can outperform SRS on variance when the quantity of interested is correlated with the inclusion probabilities (Kozak and Zieliński 2007). This is clear from the given variance formula, if we note that the π_k are constant under SRS. In this case, the term $\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)$ will be no less than $y_i - y_j$ for each pair i, j under consideration (since each $\pi_i \leq 1$). If the π_k of unequal probability sampling are approximately proportional to the y_k of interest, however, then the terms $\frac{y_i}{\pi_i}$ and $\frac{y_j}{\pi_j}$ will have a small difference since both are close to 1. The goal of unequal probability sampling is to bring the inclusion probabilities more in line with variables of interest (thus lowering the variance of our estimator) by leveraging a correlated size variable which is known for the whole population.

For some alternative intuition about this variance, consider the earlier example where we wish to estimate the total electricity consumption of a state by sampling its counties. Electricity consumption might likely be correlated with the population of a county, in which case the natural variation in simple random sampling would be amplified by the population skew and provide a highly variant estimate of total electricity consumption. That is - if we break our sample up into "small consumers" and "large consumers", the same amount of variability in which cases are sampled will cause a higher variability in the consumption among the large consumers (compared to the small), and cause larger variability in the estimate of total consumption. If we over sample the large cases, however, we would obtain a less variant estimate among the larger cases (sample variance shrinks with n) and thus a less variant estimate overall (while the variance of the small cases increases, it does not keep pace with the decrease in variance among large cases). Unequal probability sampling would give a less variant picture of the largest consumers - and thus a less variant picture of total consumption.

[14] Horvitz and Thompson developed a different estimator, but the Sen-Yates-Grundy variance estimator is generally seen to be more efficient and always produce positive estimates (Stehman and Overton 1994).

[15] In practice, it is difficult to compute the joint inclusion probabilities π_{ij} (the chance that units i and j are both included in a given sample) for an arbitrary unequal probability sampling method. Under specific sampling schemes, estimators of the Horvitz-Thompson variance have been developed that do not depend on the joint inclusion probabilities.

[16] We are (as of yet) unable to prove the Central Limit Theorem in the specific case of either minimal support sampling or random systematic sampling. Some literature discusses the applicability of the CLT for other PPS methods (Berger 1998; Overton and Stehman 1995), but nothing we found comments on these specific methods. Proceed with caution when calculating confidence intervals under these methods.

Comparing PPS sampling methods

Minimal support sampling

Given a set of inclusion probabilities, there are numerous ways to create a sample according to these probabilities. One sampling method in the class of unequal probability methods is the minimal support scheme, proposed by Deville and Tillé (1998) as a constructive application of a design that minimizes the total number of possible samples that must be considered (at most N) while still respecting the given inclusion probabilities. In the `R` programming language, this is implemented by the `sampling` package with the `UPminimalsupport()` function. The process is as follows (Tillé 2006, pp. 103):

Set $\pi(0) = \pi$. For $t = 0, 1, 2, \ldots$ until a sample is obtained,

1. Define
$$\begin{aligned} A_t &= \{k \mid \pi_k(t) = 0\}, \\ B_t &= \{k \mid \pi_k(t) = 1\}, \text{ and} \\ C_t &= \{k \mid 0 < \pi_k(t) < 1\}. \end{aligned}$$

2. Select a subset D_t of C_t such that $|D_t| = n - |C_t|$. D_t can be selected as desired; in R the function is implemented using a uniform random sample to select D_t.

3. Define
$$\alpha(t) = \min\left\{1 - \max_{k\in(C_t\setminus D_t)} \pi_k, \min_{k\in D_t} \pi_k\right\},$$
$$\pi_k^a(t) = \begin{cases} 0 \text{ if } k \in A_t \cup (C_t \setminus D_t) \\ 1 \text{ if } k \in B_t \cup D_t \end{cases}, \text{ and}$$
$$\pi_k^b(t) = \begin{cases} 0 & \text{if } k \in A_t \\ 1 & \text{if } k \in B_t \\ \frac{\pi_k(t)}{1-\alpha(t)} & \text{if } k \in (C_t \setminus D_t) \\ \frac{\pi_k(t)-\alpha(t)}{1-\alpha(t)} & \text{if } k \in D_t \end{cases}.$$

4. Select $\pi(t+1) = \begin{cases} \pi^a(t) \text{ with probability } \alpha(t) \\ \pi^b(t) \text{ with probability } 1-\alpha(t) \end{cases}$.

5. If the selected vector is a sample (that is, it only contains values from $\{0, 1\}$), stop. An observation k is included if $\pi_k(t+1) = 1$ and excluded if $\pi_k(t+1) = 0$.

This method will produce a sample of size n. A less technical explanation of the method would be:

1. Include all probability-1 cases (say there are b of these), exclude all probability-0 cases.

2. From the remaining cases (probability between 0 and 1, exclusive), draw a uniform random sample of size $n - b$.

3. Calculate a penalty term λ, which is the lesser of:

- One minus the highest inclusion probability π_k of a case k that was not included in this uniform sample
- The lowest inclusion probability π_k of a case k that was included in this uniform sample[17]

4. Accept the given sample with probability λ.
5. If rejected, rescale the probabilities for the next iteration. The effect of this rescale will alternate (unevenly) between permanently including the lowest case currently included, or permanently excluding the highest case currently excluded.

Because each iteration either permanently includes or excludes a case from the sample, the next iteration is effectively selecting from a population of size $N - 1$. Since our population is of size N, this method will consider at most N samples as its support (hence the name "minimal support"). We believe that this alternation is the source of the variation that we see empirically below. Because the method distinctly categorizes either the largest or smallest remaining case at each step, a given sample has three "boxes" when the algorithm stops running.

- The first box contains every case that was definitively tossed out until the algorithm stopped. These will be all of the smallest cases.
- The second box contains every case that was definitively tossed in until the algorithm stopped. These will be all of the largest cases.
- The third box contains the cases that were randomized on the last step of the algorithm.

Since these were sampled uniformly at random, the cases in the third box were not chosen based on their underlying size. In that way, the minimal support method can be thought of as a dynamic or randomized version of the stratification method based on size used by Kimball and Baybeck (2013): every case smaller then a is dropped, every case larger than b is included, and the cases in between are chosen uniformly at random to fill out the rest of the sample (where a and b are determined by the algorithm).

This leads to significant correlations between which cases end up being sampled in repeated draws. A small case which is not included in the sample may have been dropped at random, or it may have been dropped because it went below the lower cutoff a - in which case, every case smaller than it would also have been dropped. The more this happens (that is, the smaller the case is), the higher the correlation between its probabilty of inclusion and its neighbors' probability of inclusion. The same is true for large cases being included. If this is the case, the reason that we see occasional samples with far too many small cases is because the algorithm stopped too "early" and sampled a wide swath of the population uniformly at random, which would oversample chunks of the population that were intended to have a small inclusion probability. This phenomenon will be explored more in our simulations below.

Random Systematic Sampling

Instead of the minimal support sampling algorithm, we recommend that the random systematic method be used for populations with the same characteristics as American LEOs. The random systematic sampling algorithm is as follows:

[17] The phrase "penalty term" is a little counter intuitive here, because a higher value is a better sample. This term assesses how well our uniform sample matches the unequal inclusion probabilities. If the term is large, that means that we included the highest-probability observations and excluded the lowest-probability observations.

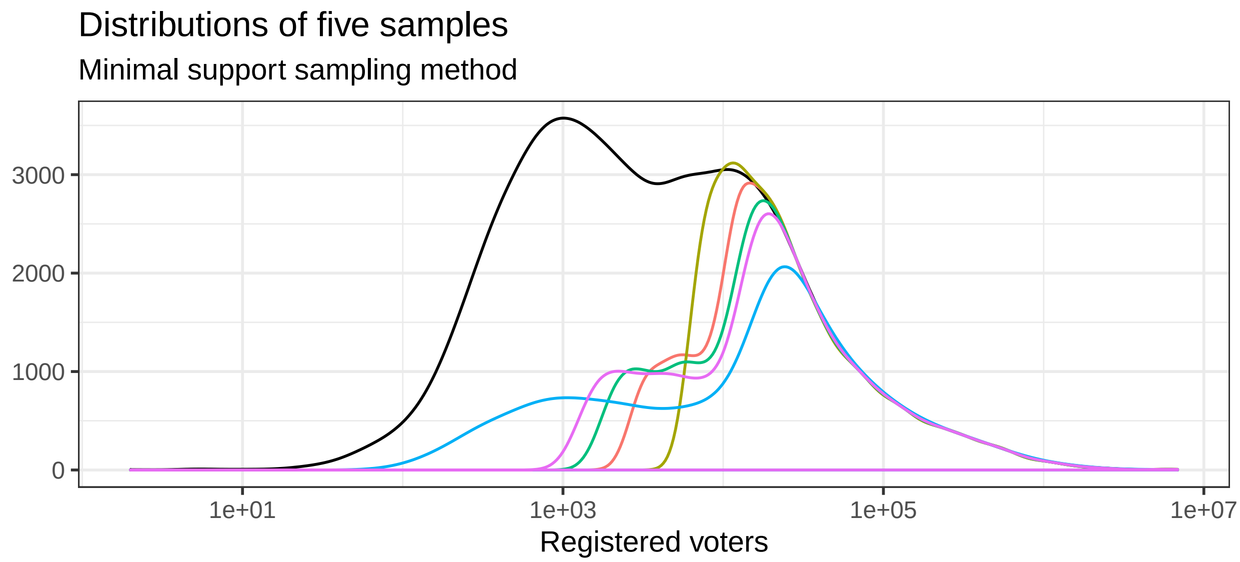

Figure 4: Distribution of 5 minimal support samples

1. Randomize the order of the units.
2. Arrange the units as line segments on a number line, where the "width" of each segment corresponds to its inclusion probability. The total length of the line segments will be the sample size n, as the inclusion probabilities sum to n in unequal probability sampling.
3. Include in the sample the cases whose segments contain any integers, or have their right endpoint on an integer (the left endpoint would be equivalent, but only one of the two should count).

This will produce a sample of size n that respects the provided inclusion probabilities. Any segment with $\pi_k = 1$ will be included, as any line segment of length 1 cannot be squeezed in between two integers without touching either one. Since the position of each segment is random, a segment of length p will have probability p of containing an integer, and thus will be sampled accordingly.

The randomization step is important to ensure independence between sampled points. Suppose there was no randomization, and the segments were ordered from shortest to longest. If there are many short segments clustered together, one segment being included would likely mean that its neighbors are not included. This correlation is something to avoid, and the randomization step does so sufficiently.

Simulations

Here we will show the empirical variation produced by the minimal support sampling scheme, as compared to other methods. Consider five samples of size $n = 3000$ conducted with the minimal support method, shown in Figure 4. The black curve above shows the density of registered voters for the entire LEO population. While any of these samples should be functionally equivalent to the others, the variation in distribution between the five is notable. This would seem to indicate that some typical properties (mean, variance, etc.) of any given sample using this method might differ widely between two functionally equivalent samples.

For comparison, consider simulations of three other common unequal probability sampling methods: Poisson Sampling, Random Systematic Sampling, and Midzuno Sampling. As we see in Figure 5, all four of

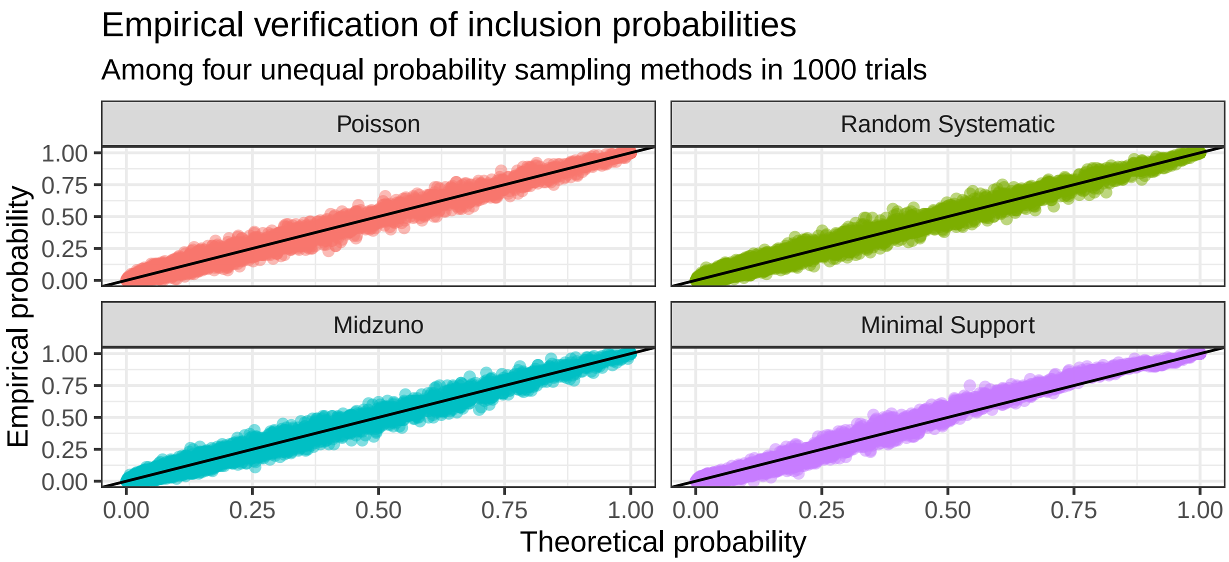

Figure 5: Empirical verification of inclusion probabilities among four PPS sampling methods

Table 2: Properties of 100 simulated sample means

Sampling method	Mean	Variance
Poisson	67037.00	204444.5420
Random Systematic	67017.35	1213.0544
Midzuno	67015.47	420.0097
Minimal Support	67176.25	768137.2175

these methods equivalently sample each jurisdiction according to its inclusion probability.

In Figure 6, we compare the distribution of registered voters in 100 samples from our population using each sampling method. While all 100 samples using the other three methods produce a similar distribution of registered voters, the minimal support method

In these three alternate sampling methods, we see much less variation in the distribution of jurisdictions sampled. To quantify this, consider the distributions of the (unweighted) sample means of total registrants from 100 simulated samples in Figure 7 and Table 2. While the three alternative methods have a very similar center of their respective distributions of sample means, the spread of the minimal support method is significantly larger than that of the Midzuno or random systematic methods.[18]

[18]The Poisson method also has large spread here, for reasons unknown. A more primary reason that we avoided Poisson sampling is that it does not guarantee an exact sample size. The Midzuno method was rejected for its complexity and computing time, as compared to the conceptually simpler and faster random systematic method.

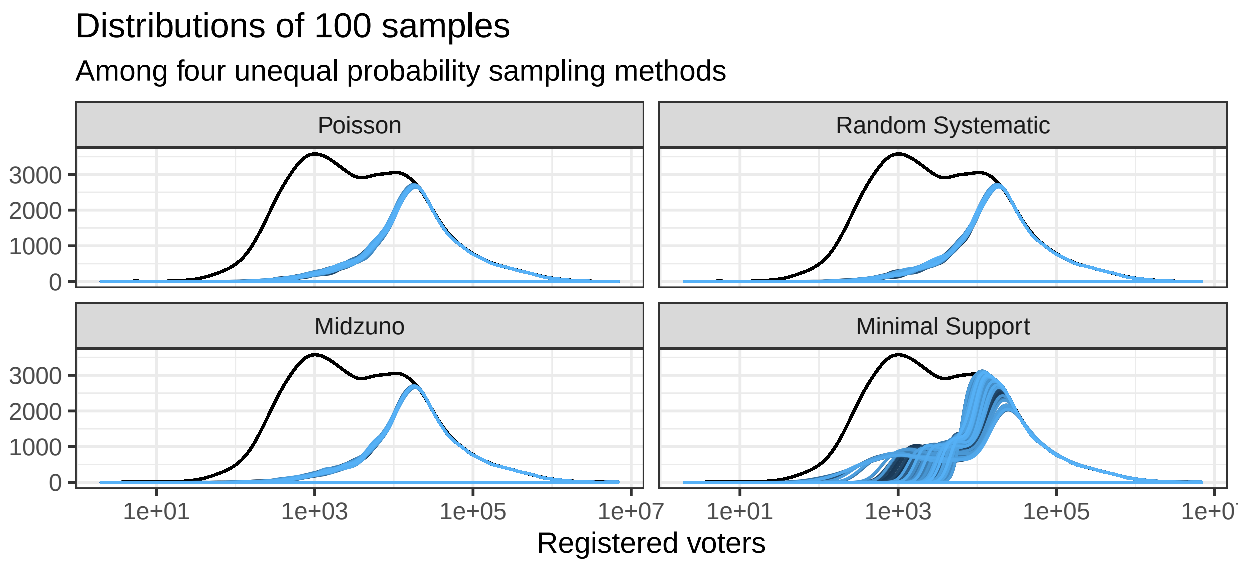

Figure 6: Distributions of 100 samples among four PPS sampling methods

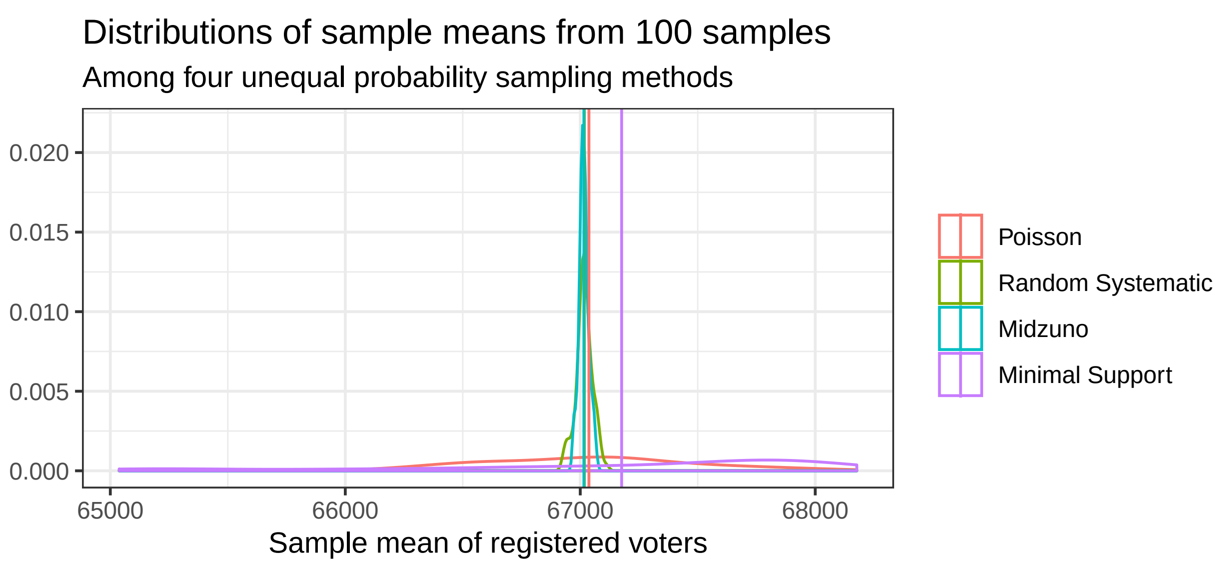

Figure 7: Distributions of sample means from 100 samples among four PPS sampling methods

The Impact of Survey Weights On Our Survey Results

We sampled proportional to size, and as a result, a larger jurisdiction is more likely to be sampled than a smaller jurisdiction. This causes our sample to over represent large jurisdictions (compared to smaller ones), and a weighting scheme will have to be employed to obtain accurate estimates of parameters such as the population mean or variance. When very small jurisdictions are included in a sample, very large weights will appear in the data. While the Horvitz-Thompson estimator does provide unbiased estimates of population parameters, large weights will increase the variance of our estimator. Additionally, the earlier issues raised about representation in regards to federalism and state laws are particularly relevant when considering sampling weights for an LEO survey.

Moving forward, we are exploring different weighting schemes with which to present our LEO survey data. In 2018, we chose to present the data unweighted and broken down into groups based on jurisdiction size, when possible. This allowed us to present the views of LEOs as a population, but still take into account the differences between LEOs in large and small jurisdictions. Because we were not weighting our final estimates or using Horvitz-Thompson estimators, the increased variance in sample mean estimates using the minimal support method was disadvantageous. We are abandoning this method, and instead used the random systematic method, which produces a much more consistent sample in a given draw.

For 2019, we have multiple options for survey weighting. Our final survey report is in the drafting stage, and we invite input on this question.

- Present unweighted results. This is complicated, because the mean is based on an arbitrary balance between large and small jurisdictions (arbitrary due to the uncertainties of response rates). Compared to the national population of LEOs, an unweighted mean under represents the smallest jurisdictions.
- Present unweighted results, but bin results into different size categories, as we did in 2018. This avoids presenting averages based on arbitrary sampling balances and conveys a clearer sense of the distribution of responses present in the LEO population.
- Present weighted results to allow generalization to the population of all LEOs. This would assign a high weight to the smallest jurisdictions (which were least likely to be sampled), and also inevitably reflects the election ecosystem of a small number of states with many small jurisdictions. The weight decreases until the cut-off point identified above, and then flat line. This focuses more on the responses of individual LEOs. Given the increase in questions about "origination stories" this year, this may be a good option. However, last year we observed significant differences on many questions between
- Present weighted results to allow generalization to the population of registered voters. This would assign a high weight to the largest jurisdictions. For example, a county with 1,000 registered voters would be weighted twice as heavily as a county with 500 registered voters (or similar). This focuses more on the impact of administrative responses to voters, but leads to some exceedingly high or low weights. In creating this weight, we would also have to consider the undersampling of small jurisdictions due to our sampling method.

All of these are complicated by nonresponse bias. If we assume that the nonresponse is truly random (not correlated with any factor that influences our results), then none of these methods are better than the others. If it is correlated with registered voters, number of employees, or election cycle scheduling, then we should attempt to adjust for non-response bias. We have some evidence that our response rates are correlated

Table 3: Breakdowns of Key 2019 LEO Characteristics

			Size (Registered Voters)		
Demographic	Overall	Weighted	$< 25{,}000$	25,000 to 250,000	$> 250{,}000$
Female	77.2%	79.8%	87.0%	68.5%	40.8%
$50,000 or more	73.8%	47.8%	62.4%	86.7%	100.0%
White	94.9%	96.6%	97.1%	92.8%	87.8%
College and above	59.5%	52.2%	52.2%	65.9%	87.8%
50 years or older	70.2%	77.0%	73.9%	65.7%	62.5%

with jurisdiction size. For example, in our 2018 survey we had more nonresponses (proportionally) from small jurisdictions than from large jurisdictions.[19] An unweighted analysis would under represent these smaller jurisdictions, tilting the balance towards large jurisdictions on a question whose answers were correlated with jurisdiction size.

For comparison, Table 3 presents some demographic information collected in our 2019 survey. The "Overall" column presents unweighted averages, the "Weighted" column presents estimates weighted according to the Horvitz-Thompson estimator (referencing the entire population of LEOs), and the final three columns are unweighted averages within three size categories. As we can see, the weights have only minimal effects on some quantities, and enormous effects on other quantities.

To illustrate, our unweighted estimate of the percentage of local election officials who are 50 years or older is reported in the lower left hand cell as 70.2%, while the weighted estimate is 77%. The difference in the estimates is caused in mathematical terms by larger survey weights applied to LEOs from the smallest jurisdictions, a larger percentage of whom report that they are 50 years old or older.

On the whole, weighted estimates will be more influenced by the smallest jurisdictions, which get comparatively larger weights, when compared to the unweighted averages. This is borne out in our preliminary results, as these variables which are all relatively correlated with size shift toward the values in the "$< 25{,}000$" column when weighted. The reason that the weighted averages sometimes go even lower than the value in this column is because the weights are calculated continously. The weighted averages show, for example, that compensation in the very smallest jurisdictions is much lower than the average compensation across the sub-25,000 category. In our survey reports, we use more detailed bins than shown here (generally opting for five categories – see Manson, Gronke, and Adona 2020), and try to be attentive to survey items that are highly correlated with jurisdiction size.

[19]In addition, in 2018, we ran a dual-mode survey (online and mail), but in 2019 ran a mail survey. The reason we ran a dual-mode survey in 2018 is that we initially began with online only, but initially had low response rates in smaller jurisdictions. Our diagnosis, reached in consultation with our advisory committees, is that cybersecurity protections put in place after 2016 made many LEOs wary of clicking on complex URLs, even when they had been pre-notified of the legitimacy of the link. A few weeks after the survey had been in the field, we produced a print version. In 2019, we had hoped that we would be able to use a single mode of administration, but this year, we had lagging responses from the largest jurisdictions, and distributed an electronic version of the instrument as a PDF. If there are future LEO surveys, we believe the dual-mode should be adoopted from the outset in order to maximize response rates and minimize non-response bias.

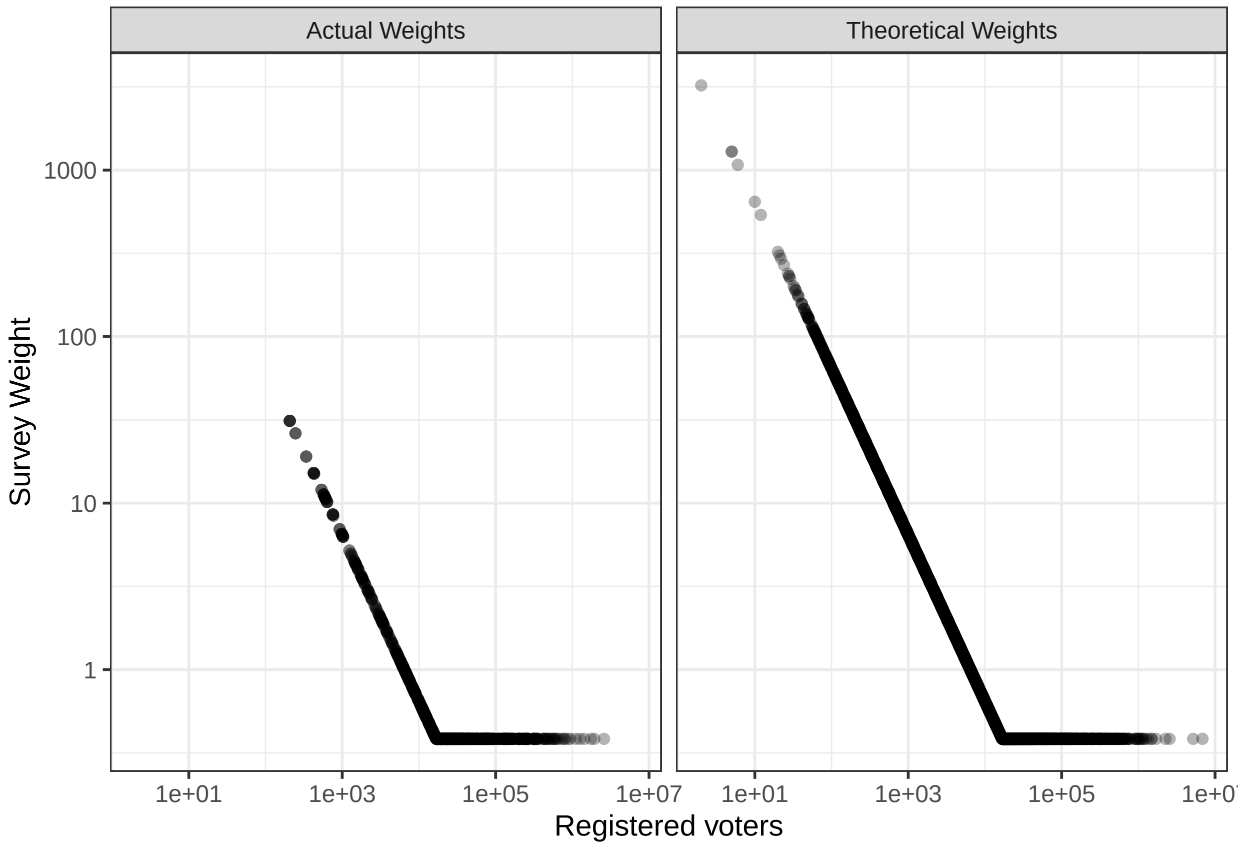

Figure 8: Horvitz-Thompson weights of LEO data

Finally, we wish to address the large weights produced by unequal probability sampling on our highly skewed dataset, and assess the practical impact of these weights on our data. The right hand panel of Figure 8 shows the weights that **would** be applied to a theoretical LEO sample. The highest of these is 3225.7, which is worrisome considering how many weights (in the medium and large jurisdictions) are less than 1.

These are only theoretical weights, however, as cases which are not sampled do not affect our sample estimates. Since the largest weights correspond to the lowest-probability cases, it is unlikely that we would sample one of these cases and have to use such a large weight. In the left hand panel of Figure 8, we report the actual weights, which reach their maximum at31.2, a much more reasonable number. This is still large compared to the many weights that are less than 1, but an appropriately-chosen sampling scheme will provide only a few cases with such large weights. While the large weights produced by this method are of great theoretical concern, they may or may not be a significant practical problem in a given sample.

Conclusion and Futher Research

Sampling a diverse and complex population like LEOs is no easy task, and significant care must be taken when surveying these stewards and reporting quantitative results about them. When surveying a population like this, we recommend using unequal probability sampling to ensure that a good spread of large and small units are considered for the sample. Compared to more manually-defined stratification methods, this can improve the weighting of individual units and reduce the variance in estimates made from the data. The skew of the population can produce some incredibly large weights, however, so a keen eye is needed to ensure that the weights sampled are reasonable.

Not all PPS sampling methods are created equal,[20] however - the minimal support method, in particular, has some anomalous behavior when applied to a population as skewed as LEOs. A simpler, faster, and more reliable method is recommended in the random systematic sampling scheme.

Our most immediate avenue for further research is to flesh out, mathematically, why the minimal support method performs so unpredictably on our LEO data. We think that the heavy skew in the data causes this undersampling problem, and perhaps more normal or unskewed data would improve the performance of this method. Some simulations may help with this as well, exploring how much skew is acceptable before the method starts to fall significantly behind. The covariance and joint inclusion probabilities between different population units seems like a good starting point for this line of inquiry, but has shown computationally prohibitive as of yet.

We are in the process of joining census data to our LEO survey data, which would give us some excellent “dependent variables” (as a function of registered voters) to test on and discover further impacts of these methods. In our simulations we were unable to explore the impacts of minimal support sampling on an unbiased, Horvitz-Thompson mean, because we did not have access to a Y variable that was known for the entire LEO population and correlated with size.

Additionally, we would be interested in exploring and quantifying the relative gains when using an unequal probability sampling method compared to a manually-stratified method like Kimball and Baybeck (2013) or simple random sampling. The claims we previously made as to the superiority of PPS are based in statistical theory, and an assessment of how much utility is gained (or lost) by using this method would be welcome.

[20]How fitting!

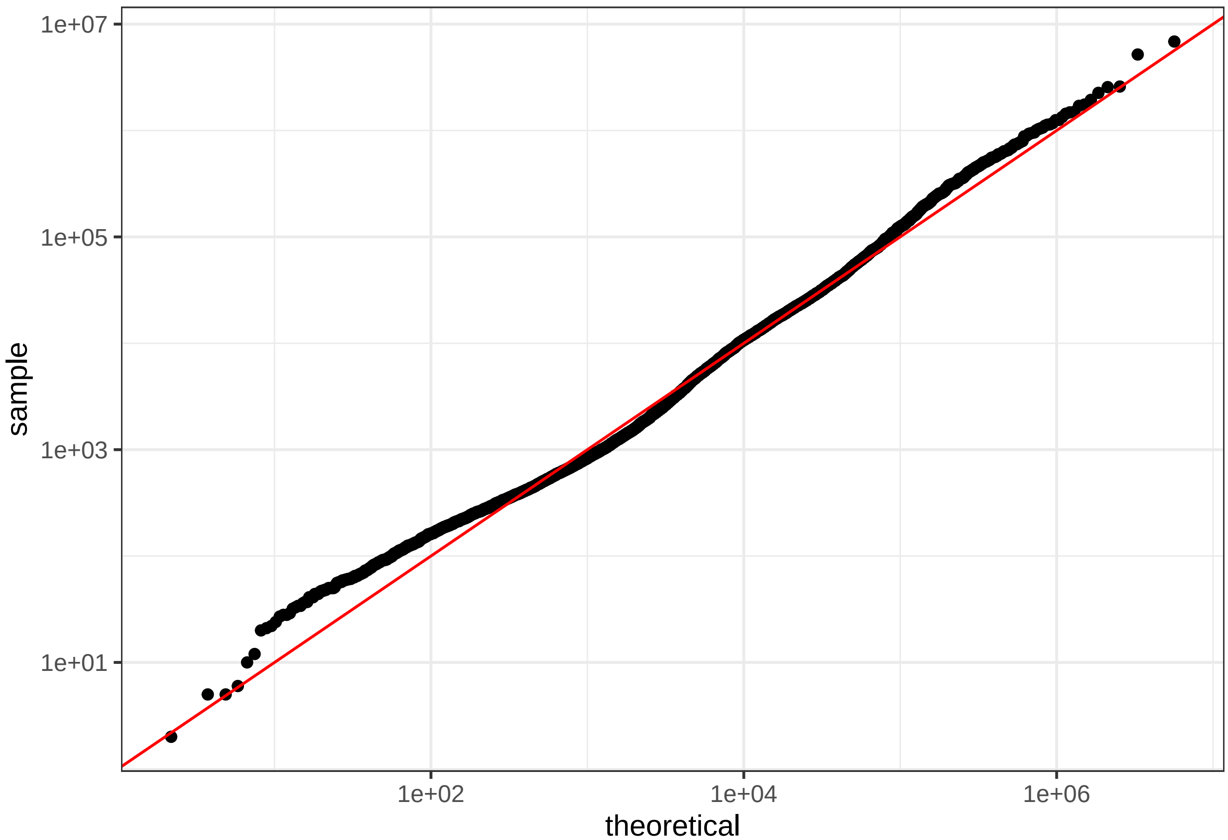

Figure 9: Quantile-quantile plot, comparing LEO population to a log-normal distribution

Appendix

Extra plots

In Figure 9, a distribution that exactly followed the theoretical log-normal distribution would trace the red line. We removed this plot from the body text but wonder if it merits inclusion.

Calculation of inclusion probabilities

Let U be our population, and let $k \in U$ be a unit of the population. Let X be a random variable that we observe for every unit of U; notably, x_k is the corresponding value for the unit k. For a sample of size 1, it is sensible to assign inclusion probabilities to each k as follows:

$$\pi_k = \frac{x_k}{\sum_{i \in U} x_i}.$$

This has the property that if one unit has a value of X twice as large as another, that unit will have a value of $\boldsymbol{\pi}$ twice as large as the other - which is desirable, if we wish to sample proportional to X.

Linear method

This can be extended to the case where $n > 1$, using the following algorithm. In `R`, this algorithm is implemented by Tillé and Matei's `sampling` package with the `inclusionprobabilities()` function, what we refer to in the rest of this paper as the **minimal support** method.

1. For all $k \in U$ such that $x_k \leq 0$, set $\pi_k = 0$. Define $U_0 := \{k \in U | x_k > 0\}$.
2. For all $k \in U_0$, compute
$$p_0(k) = \frac{n_0 x_k}{\sum_{l \in U_0} x_l}.$$
3. For the subset $U_0^+ = \{k \in U_0 \mid p_0(k) >= 1\}$, set $\pi_k := 1$.
4. Define $U_1 = \{k \in U_0 \mid p_1(k) < 1\}$, and $n_1 = n_0 - \left|U_0^+\right|$.
5. For all $k \in U_1$, compute
$$p_1(k) = \frac{n_1 x_k}{\sum_{l \in U_1} x_l}.$$
6. For the subset $U_1^+ = \{k \in U_1 \mid p_1(k) >= 1\}$, set $\pi_k := 1$.
7. Continue until i satisfies $p_i(k) < 1 \; \forall \; k \in U_i$. For these k, set $\pi_k = p_i(k)$.

   In other words:

1. For cases where x_k is not positive, set π_k to 0.
2. If x_k is greater than the sum of X divided by the sample size n, set π_k to 1.
3. For the remaining cases (those not yet given a probability; call these U_1), figure out how many are left to sample. This will be n minus however many were previously set to 1. Call this number n_1.
4. A second time, if x_k is greater than the sum of X (summed over our smaller U_1) divided by the (smaller) sample size n_1, set pi_k to 1.
5. Keep assigning probability-one cases until none of the unassigned x_k are larger than the current metric of $\sum_{U_z} x_i$ over n_z. For each of these, set $\pi_k = \frac{n_z x_k}{\sum_{i \in U_z} x_i}$.

These inclusion probabilities $\boldsymbol{\pi}$ are in $[0,1]$, and proportional to the value of X (until reaching a ceiling of 1). They can equivalently be expressed as

$$\pi_k = \min\left\{1, h^{-1}(n)\frac{x_k}{\sum_{k \in U} x_k}\right\} \tag{6}$$

for each $k \in U_0$, where

$$h(z) = \sum_{k \in U_0} \min\left(z\frac{x_k}{t_x}, 1\right). \tag{7}$$

It is a consequence of Equation (6) that these inclusion probabilities depend on every value x_k in the population, that is the specific distribution of x_k. Individual changes can alter the values for $h(z)$ and t_x, and can also impact which cases get set to $\pi_k = 1$ in each step of the algorithm.

Sequential method

An alternative, and perhaps more straightforward, formulation of the inclusion probabilities would be the result of a sequential sampling design as follows:

1. Sample 1 unit, proportional to size. That is, if a unit has size 5 and the overall size of the population is 100, that unit has a 5% chance of being selected.
2. Remove that unit from the population. If the unit with size 5 is selected, our population's total size is now 95.
3. Sample a second unit, proportional to size in the new (smaller) population. Now, a unit with size 4 would have a $4/95 \approx 4.2\%$ chance of being sampled.
4. Remove that unit from the population. If the unit with size 4 is selected, the total size is now 91.
5. Continue until n units are selected.

At each step in this method, we are selecting a single unit, with probability proportional to its size among the remaining cases. While obtaining a closed-form sampling probability for each unit is difficult for large n, we can simulate many samples using this method to approximate the sampling probabilities for a given case.

Figure 10 compares the formulaic inclusion probabilities obtained from the linear method to the simulated inclusion probabilities obtained from the sequential method. We simulated 10,000 samples for the sequential method, to get some convergence in the line connecting each point. Since there is not a formulaic expression for a given inclusion probability, there is no guarantee that two jurisdictions with the same size will get the same inclusion probability under the simulation. By increasing the number of trials, the inclusion probabilities for a given size are less variant and the line in the plot becomes more understandable and accurate.] as applied to our LEO data. The x-axis of the plot is truncated at 75,000 so the differences in methods are visible at scale.

Comparison of two methods

Now, consider the inclusion probabilities from the linear method that we outlined earlier. Up until the cutoff point, each π_k is proportional to x_k, and after the cutoff point each $\pi_k = 1$. Consider pairs $i, j \in U$ with $i < j$,[21] as in the summation of the variance of the Horvitz-Thompson mean estimator. Suppose one of the cases has $\pi_k = 1$, without loss of generality call this case j (the math is equivalent for i). Since j is always sampled, then $\pi_{ij} = \pi_i$ and the term in the variance summation corresponding to i, j is

$$(\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = (\pi_i \cdot 1 - \pi_i) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = 0.$$

So for any pair i, j where either $\pi_i = 1$ or $\pi_j = 1$, the contribution to the variance sum is 0. When both π_i and π_j are less than one, however, they are proportional to x_i and x_j, respectively. For a variable Y that is correlated to X, this will make the y_i and y_j proportional to π_i and π_j (in the aggregate, at least), and reduce the contribution to the variance sum. Combined with the zero contribution from cases included

[21] While $i < j$ may not make sense in all contexts, the notation indicates that we wish to sum over all pairs of elements i, j where $i \neq j$, and exclude the reverse j, i of each of these pairs. Since the expression inside the sum does not depend on which case is i or j, we could equivalently sum over the i, j and j, i and divide the result by two.

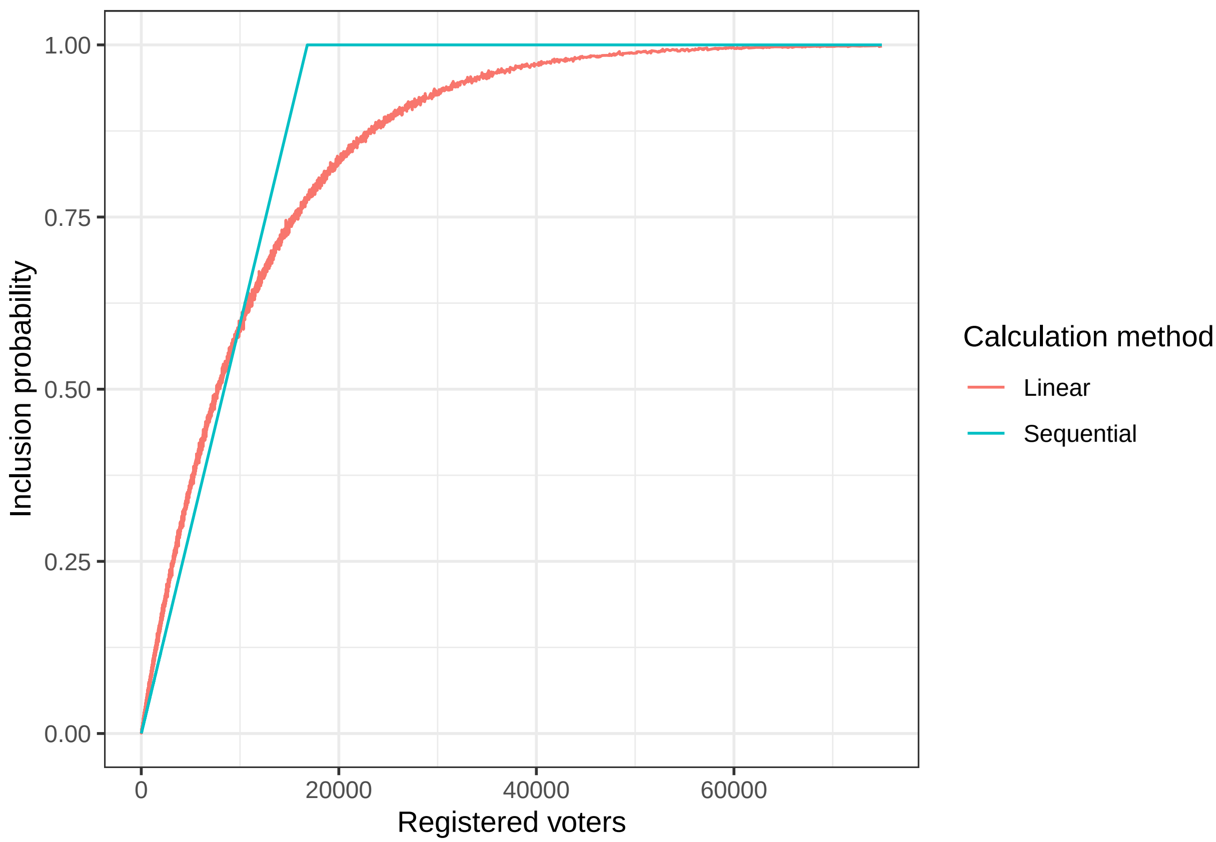

Figure 10: Comparison of methods for obtaining inclusion probabilities

with certainty, we see that this linear formulation of inclusion probabilities minimizes the variance of the Horvitz-Thompson mean estimator for variables correlated with size.

References

Adona, Natalie, Paul Gronke, Paul Manson, and Sarah Cole. 2019. "Stewards of Democracy: The Views of American Local Election Officials." Washington, D.C.: Democracy Fund. https://www.democracyfund.org/publications/stewards-of-democracy/.

Berger, Yves G. 1998. "Rate of Convergence to Normal Distribution for the Horvitz-Thompson Estimator." *Journal of Statistical Planning and Inference* 67 (2): 209–26. https://doi.org/10.1016/S0378-3758(97)00107-9.

Burns, Nancy. 1994. *The Formation of American Local Governments: Private Values in Public Institutions.* Oxford University Press.

DesRoches, David. 2011. "Establishment Survey." Edited by Paul Lavrakas. *Encyclopedia of Survey Research Methods.* 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. https://doi.org/10.4135/9781412963947.

Fischer, Eric A., and Kevin J. Coleman. 2005. "What Do Local Election Officials Think About Election Reform?" Congressional Research Service.

Government Accountability Office. 2018. "Elections: Observations on Voting Equipment Use and Replacement." GAO-18-294. Washington DC.

Gronke, Paul, and Evan Crawford. 2019. "A Locals Guide to the State of Election Administration." In *Southern Political Science Association.* Austin, TX.

Hale, Kathleen, Robert Montjoy, and Mitchell Brown. 2015. *Administering Elections: How American Elections Work.* New York: Palgrave Macmillan.

Horvitz, D. G., and D. J. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47 (260): 663–85. https://doi.org/10.2307/2280784.

"How Local Election Officials View Election Reform: Results of Three National Surveys." 2011. Washington, DC: Congressional Research Service. https://www.everycrsreport.com/files/20110304_R41667_0906f3151353ce165dccfc85deda70203f6a245c.pdf.

Huefner, Steven F, Daniel P Tokaji, Edward B Foley, Nathan A Cemenska, Moritz College of Law, and Election Law @ Moritz. 2007. *From Registration to Recounts: The Election Ecosystems of Five Midwestern States.* Columbus, Ohio: Ohio State University Michael E. Moritz College of Law.

Kimball, David, and Brady Baybeck. 2013. "Are All Jurisdictions Equal? Size Disparity in Election Administration." *Election Law Journal* 12 (2). https://doi.org/10.1089/elj.2012.0174.

Kimball, David C, Jennifer Collins-Foley, Connie Schmidt, and Chere' Maxwell. 2010. "SURVEY OF POLL WORKER RECRUITMENT, TRAINING, AND EVALUATION PRACTICES BY LOCAL ELECTION OFFICIALS." http://www.umsl.edu/~kimballd/report_june112010.pdf.

Kozak, Marcin, and Andrzej Zieliński. 2007. "Comparison of Efficiency of Stratified and Unequal Probability Sampling." *Communications in Statistics - Simulation and Computation* 36 (4): 807–16. https://doi.org/10.1080/03610910701419695.

Kropf, Martha, and JoEllen V. Pope. 2019. "Election Costs: A Study of North Carolina." In *The Future of Election Administration*, 185–200. New York: Palgrave Macmillan.

Lavrakas, Paul, and Michael P. Cohen, eds. 2011. "Auxiliary Variable." *Encyclopedia of Survey Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. https://doi.org/10.4135/9781412963947.

Moynihan, Donald P., and Carol L. Silva. 2008. "The Administrators of Democracy: A Research Note on Local Election Officials." *Public Administration Review* 68 (5): 816–27. https://doi.org/10.1111/j.1540-6210.2008.00923.x.

Overton, W. Scott, and Stephen V. Stehman. 1995. "The Horvitz-Thompson Theorem as a Unifying Perspective for Probability Sampling: With Examples from Natural Resource Sampling." *The American Statistician* 49 (3): 261–68. https://doi.org/10.2307/2684196.

"Sampling Techniques." n.d. Food and Agriculture Organization of the United Nations. Accessed October 17, 2019. http://www.fao.org/3/X6831E/X6831E12.htm.

Sen, A. R. 1953. "On the Estimate of the Variance in Sampling with Varying Probabilities." *Journal of the Indian Society of Agricultural Statistics* 5: 119–27.

Stehman, Stephen V., and W. Scott Overton. 1994. "Comparison of Variance Estimators of the Horvitz-Thompson Estimator for Randomized Variable Probability Systematic Sampling." *Journal of the American Statistical Association* 89 (425): 30–43. https://doi.org/10.2307/2291198.

Sunter, A. B. 1977. "List Sequential Sampling with Equal or Unequal Probabilities Without Replacement." *Applied Statistics* 26 (3): 261. https://doi.org/10.2307/2346966.

"Survey of Local Election Officials." 2019. Early Voting Information Center. 2019. https://blogs.reed.edu/earlyvoting/leo-survey/.

Tillé, Yves. 2010. "TEACHING SURVEY SAMPLING WITH THE 'SAMPLING' R PACKAGE."

———. 2016. "Sampling: Survey Sampling R Package Version 2.8."

Yates, F., and P. M. Grundy. 1953. "Selection Without Replacement from Within Strata with Probability Proportional to Size." *Journal of the Royal Statistical Society*, Methodological, 15 (2): 253–61. https://www.jstor.org/stable/2983772.