Cut-off and coordinated survey sampling

Kira Coder Gylling
Örebro University and Statistics Sweden

September 22, 2023

Abstract

The problem of estimation from a sample is well known and its solutions are the foundation for the sample surveys conducted at e.g. national statistical institutes. This dissertation looks at the estimation problem from the perspective of business survey methodology. The focus is on two types of survey sampling: Cut-off sampling and coordinated sampling. We introduce and evaluate new method for estimation in cut-off sampling as well as a tool for drawing coordinated samples in the R programming language.

Acknowledgements

I would like to thank my advisors for patient guidance and support throughout this process, my employer for equally patient financing despite delays, and my lovely wife for the particularly patient task of always being there for a stressed academic.

1 Introduction

The problem of estimating a population total from a subset of the population has been rigorously studied for at least about a century (Neyman, 1934). A central issue is that of representativity. If you just send out a questionnaire without prior knowledge of the population then you cannot know which and how large proportion of the population each respondent represents.

This can be alleviated by drawing a sample from a sampling frame – an enumerated list of all the units in the population of interest – and weighting their responses with the inverse of their respective probability of sample inclusion before adding. It has been shown by Horvitz and Thompson (1952) that under various conditions this leads to an unbiased estimate of the total. Moreover, the estimates can be made more precise and accurate by stratifying the frame before sampling, i.e. to divide it up into a partition of subsets and taking a subsample in each subset.

For business surveys in particular there are certain conflicting considerations to be taken into account. On the one hand, bigger businesses typically contribute more to e.g. economic estimates than smaller ones so we want as many big businesses as possible in the surveys. On the other hand, we might not want to include the same businesses in several surveys – or perhaps we do.

When the size of the business is relevant then it is not uncommon to use cut-off sampling (e.g. Benedetti et al. 2010) – to exclude businesses that are smaller than a certain threshold from the frame and only sample among the bigger ones. Obviously this inserts bias into the estimates. One way of trying to reduce this bias is to use model estimates for the smaller businesses. Whether this actually reduces the bias one cannot really tell.

Furthermore, when more than one survey or more than one instance of the same survey is considered, then coordinated sampling can help with either spreading out the response burden by reducing the overlap between samples, or reducing variance by increasing the overlap. The former case is called negative coordination, and the latter positive coordination.

In the following papers we first present a novel method of making cut-off samples representative by including old information on businesses that previously were excluded from the frame. Next we present a tool for drawing coordinated samples using permanent random numbers.

2 Survey sampling

2.1 Background

Survey sampling is used by statistical agencies and other institutes that need accurate information about a population without having to ask everyone. Instead a sampling frame – a list or table of all units in the population – is used, in which each unit k is given probability π_k of being included in the sample, where $\pi_k \in (0, 1)\ \forall k$ and where

$$\sum_{k \in U} \pi_k = n$$

for population U, population size N and sample size n. Next a study variable y is collected for each unit in the sample, and the population total Y is estimated as

$$\hat{Y} = \sum_{k \in s} \frac{y_k}{\pi_k}$$

for sample s. For simple random sampling, where each unit has the same inclusion probability, this simplifies to

$$\hat{Y} = \frac{N}{n} \sum_{k \in s} y_k$$

It has been shown (Horvitz and Thompson, 1952) that this is an unbiased estimation procedure.

If an correlating auxiliary variable x is available then the ratio estimator (Royall and Cumberland, 1981) can be used instead,

$$\hat{Y} = \frac{X_U}{X_s} \sum_{k \in s} y_k$$

where

$$X_U = \sum_{k \in U} x_k$$

and

$$X_s = \sum_{k \in s} x_k$$

For simple random sampling it is easy to see that this is the Horvitz-Thompson estimator augmented with more information to replace the less specific factor N/n.

A more powerful estimation procedure that can take an arbitrary number of auxiliary variables into account is the general regression estimator (Deville and Särndal, 1992), which is a form of calibration estimator. The fundamental idea of calibration is to use weights that are calculated such that they are as close as possible to the design weights and also satisfy a number of constraints called calibration equations. Deriving these weights is a constrained minimization problem which involves Lagrange multipliers and which is outside the scope of this dissertation.

However, one important area of application for calibration is surveys with non-response (Lundström and Särndal, 1999). Calibrating against known population totals for auxiliary variables can give us weights that take into account the presence of non-response and thus we can have a more accurate estimate.

2.2 Cut-off sampling

Consider a sampling frame where each unit contains a size measure X, for instance number of employees in the case of business surveys. We define a stratified design by taking three strata along X with sample sizes and inclusion probabilities that vary greatly between strata. Below we list an example of such a stratification.

$$\begin{cases} A: & X \in [K, \infty), & n_A = N_A, & \pi_k = 1 \\ B: & X \in [k, K), & n_B < N_B, & \pi_k \in (0, 1) \\ C: & X \in [0, k_1), & n_C = 0, & \pi_k = 0 \end{cases}$$

Here, all large enough units, i.e. ones with $X \geq K$, are put into a take-all stratum A and are thus included with probability 1. Between K and k there is a take-some stratum B, where some but not all units are sampled. k is a cut-off threshold, where any unit smaller than k is given inclusion probability 0 and is in effect excluded from the frame. To simplify the analysis we put these into their own stratum C, a situation that Baillargeon and Rivest (2009) calls a "take-none" stratum.

For the sake of simplicity, we assume that we draw a simple random sample from this population. Then the values of π_k are 1, n_B/N_B, and 0. Collecting values of the variable of interest y for the sampled units, we get an unbiased estimate of the population total Y_{above} for the population above the cut-off threshold as

$$\hat{Y}_{above} = \sum_{k \in A} y_k + \sum_{k \in B} \frac{N_B}{n_B} y_k.$$

It is now easy to see that, since no information from C is included, using $\hat{Y}_{above}$ as an estimate of the total for the entire population would lead to bias – particularly an underestimate, assuming y is non-negative. However, there are situations where we want this kind of cut-off sampling design, e.g. to minimize cost or response burden. We therefore need inference methods to compensate for this missing information.

Benedetti et al. (2010) use auxiliary information, known for the entire population, to model how large proportion of the target population total is missing due to the take-none stratum. This is then expressed as a fraction δ, and the total estimate is adjusted by a factor $(1 + \delta)$. Knaub (2007) reviews the use of a ratio estimator with one auxiliary variable and regression predictions replacing y for the take-none stratum. Elisson and Elvers (2001) looks at two regression estimators and conclude that, while cut-off sampling can be useful, it leads to a difficult estimation problem, and furthermore that care must be taken when choosing a size variable.

When read critically, the aforementioned papers all seem to make a few assumptions:

1. Correcting for the excluded information is done during estimation.
2. Information cannot be collected from below the threshold.
3. Inclusion probabilities are identically zero and cannot be used.

These assumptions are reasonable but yield complicated estimators with an unknowable bias. An alternative approach is to find a way to make the cut-off sample representative of the excluded population and directly estimate inclusion probabilities to use in estimation. Prior to this work, no research has looked into this approach.

2.3 Coordinated sampling

Coordinated sampling is a useful tool. Suppose we have either more than one survey, or more than one occasion of the same survey. In either case, depending on what quality priorities we make, we might want to maximize compatibility, spread out response burden, minimize variance, or something else. Each of these amount to ways to control the overlap between the samples, either by maximizing it – positive coordination – or by minimizing it – negative coordination.

To see how two samples can be coordinated automatically and with precision, consider a standard procedure for drawing, for instance, stratified simple random samples or stratified Pareto probability-proportional-to-size (πps) samples. Without loss of generality, we assume that the number h of strata equals 1, if not then the routines are repeated for each stratum.

For simple random sampling with sample size n and population size N, we generate N random numbers U_k that follow a continuous uniform distribution between 0 and 1, attach them to the sampling frame such that each unit is associated with a random number, and sort the frame increasingly along U_k. Then the n units with the lowest U_k – the top n elements of the sorted frame – are sampled.

For Pareto πps sampling (Rosén, 1997) with similar notation for sample size and population size, we first define a measure x_k of the size of a unit k, with which we calculate the approximate inclusion probabilities

$$\lambda_k = n\frac{x_k}{\sum x_k}$$

where the sum is taken over all elements in the relevant stratum. Next we generate the same kind of random numbers U_k as in simple random sampling. However, instead of sorting directly on U_k we calculate a rank Q_k that incorporates the approximate inclusion probabilities λ:

$$Q_k = \frac{U_k(1-\lambda_k)}{\lambda_k(1-U_k)}$$

We sort the frame increasingly along Q_k and the n units with the lowest Q_k are sampled. The factor $(1-\lambda_k)/\lambda_k$ means that units with higher λ_k are sampled more often than ones with lower λ_k, and the factor $U_k/(1-U_k)$ means that this difference is probabilistic.

Notice how in each of these sampling routines, U_k is the only source of randomness. Thus an intuitive way of achieving reproducibility is to pre-generate the U_k's before the sampling routine starts, attach them to the sampling frame, and use these pre-generated – or permanent – random numbers instead of letting the routine generate them. Then we can reuse them and get the sample twice. Furthermore, by choosing to sample from e.g. 0.4 increasingly or 1.0 decreasingly instead of 0.0 increasingly, we can control how many units from the first sample are sampled the second time. Thus we have a method for sample coordination using permanent random numbers.

Since this is by its nature a computationally intensive process there is a need for software packages that can support it. So far, there exists such a software package for SAS, called DraUrval and developed at and for Statistics Sweden. Sadly all documentation for it is currently internal to Statistics Sweden. Prior to the work around which this dissertation is centered, no such package existed in the R programming language.

3 Summary of papers

3.1 Quasi Randomization and Cut-Off Samples

With some data in missing by design, we start from the assumption that cut-off sampling can be viewed as designed non-response. This opens up for the possibility of applying methods that were previously used for non-response to this type of sampling. We investigate the so-called quasi-randomization approach to calculating response probabilities (Oh and Scheuren, 1983), and apply it to cut-off samples.

We calculate inclusion probabilities based on historical values of the size variable – the variable along which the cut was made – and start from the assumption that a unit that was not excluded at year t but was excluded at the previous year $t-1$ could be representative of its size class at $t-1$. By collecting one-year lagged data on such units at year t we should be able to calculate an accurate estimate of the excluded population for year $t-1$.

This is in fact the case, which we show mathematically as well as empirically via a fictional register-based cut-off sample survey. We use number of employees as size variable X, excluding all below 10, and turnover as study variable y. Treating the subset of the population with $X_t \geq 10$ and $X_{t-1} < 10$ as a Poisson sample with equal probabilities, we find that for each domain defined

by $X_{t-1} = 1, \ldots, 9$, the sample mean is sufficiently close to the true mean to conclude that the method is sound. At $X = 0$ the method does not perform as well due to selection bias: Businesses that grow from zero to ten employees from one year to the next are not representative of the vast majority of businesses with zero employees – particularly in terms of turnover.

We conclude that the method is potentially useful but that care should be taken in modelling the inclusion probabilities, to avoid introducing bias in estimates.

3.2 prnsamplr: Permanent Random Number Sampling in R

As is mentioned in Section 2.3 there is a need for software tools that can draw coordinated samples using permanent random numbers (PRN's). In this paper we present an R package that has been developed specifically for this purpose. It can be used for both positive and negative coordination in stratified simple random sampling as well as stratified Pareto πps sampling, in the manner described in Section 2.3, and we show its performance on an artificial dataset.

We find that the tool behaves as expected: Drawing samples from two nearby points and in the same direction leads to a relatively high overlap, while drawing samples from two farther-away points or in opposite directions gives a lower overlap. Drawing samples from the same point and in the same direction gives 100% overlap, showing that there is no randomness introduced by the tool.

4 Discussion

There are two parts of statistical inference: Sampling and estimation. The former deals with determining from whom we gather information, and the latter deals with how we can draw conclusions about the population given this information. Sampling influences estimation by limiting which types of estimation we can make given the sampling design, and conversely, estimation influences sampling by limiting which sampling design we can use given the type of estimation we want to make. In the statistics production process there are several other steps before, between, and after sampling and estimation, but these are the two inference steps.

If we want to be rigorous, then these steps cannot be performed in any ad-hoc manner we want. Instead we need clearly defined procedures that are based on scientifically sound methods and aided by useful tools. In these papers we look at methods and tools for the sampling step of the statistics production process.

In the first paper we look at a certain type of sampling that does not sample from the entire population, namely cut-off sampling, where only the largest units in the population are sampled. Cut-off sampling is a cost-effective method but one that is prone to bias. This paper suggests reducing this bias by collecting past data on units that only recently grew past the threshold for being sampled, and using such data in order to make inferences on the population that is below the threshold. This is a new contribution to the field of cut-off sampling. However, to further develop the method, it is important to test it on survey data rather than register data. This is to see how well it performs compared to methods based on, for instance, model estimation or taking a supplementary sample below the threshold. Then we can develop theoretically sound models for estimating the probability that a business moves over the threshold from one year to the next.

The second paper approaches the need for useful tools in a computer-intensive field that lacks them, namely coordinated sampling. We introduce a new software tool, written in R and distributed as a package on CRAN (Coder Gylling, 2023), that makes it possible to use R for drawing stratified

simple random samples as well as stratified Pareto πps samples, with or without coordination. There are several R packages with similar functionality, primarily sampling (Tillé and Matei, 2021), survey (Lumley, 2023), and pps (Gambino, 2021). However, this is the first R package supporting permanent random number sampling that is available on CRAN. It is also the first R package on CRAN that supports stratified Pareto πps sampling. There are however several πps routines that the package cannot yet handle, for instance systematic πps sampling. Work has also not yet been done at making this tool fully compatible with the data structures and syntax of such packages as survey or dplyr. The reader is more than welcome to contribute, since the package is open source.

References

S. Baillargeon and L.-P. Rivest. A general algorithm for univariate stratification. *International Statistical Review / Revue Internationale de Statistique*, 77(3):331–344, 2009. ISSN 03067734, 17515823. URL `http://www.jstor.org/stable/27919760`.

R. Benedetti, M. Bee, and G. Espa. A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4):651–671, 2010.

K. Coder Gylling. *prnsamplr: Permanent Random Number Sampling*, 2023. URL `https://CRAN.R-project.org/package=prnsamplr`. R package version 0.3.0.

J.-C. Deville and C.-E. Särndal. Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382, 1992. ISSN 01621459. URL `http://www.jstor.org/stable/2290268`.

H. Elisson and E. Elvers. Cut-off sampling and estimation. In *Proceedings of Statistics Canada Symposium*, 2001.

J. G. Gambino. *pps: PPS Sampling*, 2021. URL `https://CRAN.R-project.org/package=pps`. R package version 1.0.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459. URL `http://www.jstor.org/stable/2280784`.

J. R. Knaub. Cutoff sampling and inference. *InterStat*, 2007.

T. Lumley. survey: analysis of complex survey samples, 2023. R package version 4.2.

S. Lundström and C.-E. Särndal. Calibration as a standard method for treatment of nonresponse. *Journal of official statistics*, 15(2):305, 1999.

J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625, 1934. ISSN 09528385. URL `http://www.jstor.org/stable/2342192`.

H. L. Oh and F. J. Scheuren. Weighting adjustment for unit nonresponse. In W. G. Madow, I. Olkin, and D. B. Rubin, editors, *Incomplete Data in Sample Surveys*, volume 2, pages 143–184. Academic Press, New York, 1983.

B. Rosén. On sampling with probability proportional to size. *Journal of statistical planning and inference*, 62(2):159–191, 1997.

R. M. Royall and W. G. Cumberland. An empirical study of the ratio estimator and estimators of its variance. *Journal of the American statistical Association*, 76(373):66–77, 1981.

Y. Tillé and A. Matei. *sampling: Survey Sampling*, 2021. URL `https://CRAN.R-project.org/package=sampling`. R package version 2.9.

Quasi Randomization and Cut-Off Samples

Kira Coder Gylling and Thomas Laitila
Örebro University and Statistics Sweden

Abstract

Cut-off samples, where many small units are explicitly removed from the sampling frame, are common in business statistics settings where the study variable is assumed to correlate with the size variable. The problem that then arises is how to estimate the population total for the subset of the population that is cut off from the frame. In this article we suggest a method based on the insight that units that are included in the sample one year can be representative of the population outside of the sample the previous year. We find that under the right conditions this representative can be captured and the population total can be estimated with approximately unbiased estimators.

Keywords sampling, business surveys, modelling

1 Background

Business populations tend to be skewed. There are a few very large businesses that contribute a lot to economic statistics, and many small ones that individually do not contribute much. In such circumstances, national statistical institutes might decide to base their business surveys on cut-off samples: samples drawn from a frame from which each business below a certain size has been excluded. This way one gets a certain bias but in return gets lower variance, cost, or response burden.

Despite their usefulness in official statistics production, not a lot has been published about cut-off samples. Most papers, such as Elisson and Elvers (2001), Benedetti et al. (2010), and Hwang and Shin (2013), aim at designing estimators. For a general introduction to the subject, see Knaub (2008). Haziza et al. (2010) introduced an interesting setting where auxiliary information is used for both the estimation step and for creating pseudo-weights for estimating the population total for the sub-population that was excluded from the sampling frame.

This paper uses a similar setting as the latter one. However, we do not view the population as static. We rather base the pseudo-weights on a Quasi Randomization interpretation. Probabilities are defined in terms of how large the units were during the year before the sample was drawn, and use their values of the variable of interest at this previous year in order to estimate a lagged population total. This allows us to use simpler estimators than in Haziza et al. (2010), and under the right conditions we can define approximately unbiased estimators.

In Sections 2 through 4 we present the theoretical underpinnings of the method. Section 5 includes an example of the risks of selection bias.

2 Quasi Randomization and Cut-Off Samples

2.1 Quasi Randomization

The approach of treating non-response as an outcome of a random trial is by Oh and Scheuren (1983) named "Quasi Randomization" (QR). Here the subset of respondents from the sample is assumed generated by randomization while the sampling design is unknown, as are the individual units' response probabilities.

Adapting the QR-approach paves the way for cheaper and faster statistics compared with re-sampling of non-respondents. The sampling design for the original sample combined with the design for the response set can then be interpreted as a two-stage sampling design. If the second stage sampling design is invariant to the first, the probability of the k:th population unit being included in the response set is $\pi_k = \pi_{1k}\theta_k$, where π_{1k} is the probability of inclusion in the first stage sample $s \subset U$ (U denotes the population set), and θ_k is the inclusion probability in the second stage.

The probability of sample inclusion of a unit in the first stage is defined by the design and is known. The inclusion probabilities in the second stage are not known, however. The idea is instead to formulate and estimate a probability model for sample inclusion, usually under an independence assumption, i.e. $P(r \supset \{k,l\} \mid \{k,l\} \subset s) = \theta_k\theta_l$ denoting the response set with r.

Little (1986) suggests estimation of a logit model with auxiliary variables known for all sampled units. Then estimated probabilities are used for dividing sample units into adjustment cells and within cells constant response probabilities are calculated.

Let $\hat{\theta}_k$ denote estimated inclusion probabilities in the second stage. Then the direct weighting estimator is defined as

$$\hat{Y} = \sum_r \frac{1}{\pi_{1k}\hat{\theta}_k} y_k \tag{1}$$

where y_k is the study variable and Y the population total.

2.2 Cut-Off sampling

A similar QR-approach is here suggested for cut-off sampling. A variable x_k known for all units in the population is used to divide the population into two sets: $U_1 = \{k : x_k \geq c, k \in U\}$ and $U_0 = U - U_1$. A sample s is drawn from the set U_1 while no units are drawn from U_0, they are "cut off" from the population studied.

Having the sample units from U_1 and their responses, inference may either be on the cut-off population U_1 or the whole population U. These combinations of sampling and inference is by Knaub (2008) defined as methods 1 and 2, respectively. Method 1 is within the standard randomization theory and poses no problem with respect to how inference is to be made. Method 2 on the other hand involves extrapolation of results valid for U_1 to the units cut off in U_0. This can only be made by adding assumptions on how units in the two subpopulations relates to each other. An example is to use the sample from U_1 to estimate an assumed regression model valid for the whole population U. Estimates of study variables are then obtained by predictions for units in U_0.

Treating x_k as the outcome of a random trial cut-off sampling seems to mimic the QR-approach to survey non-response, but in reversed order. The QR selection is in the first stage while the traditional randomization sampling design is the second stage. In general the second stage will not be invariant to the first stage meaning the sampling follows a two-phase design. For a given

U the probability a unit is in the cut-off population U_1 is $\theta_k = Pr(U_1 \ni k) = Pr(x_k \geq c)$, where c is the cut-off threshold. Given it is included in U_1 its sample inclusion probability is $\pi_{k|1} = Pr(s \ni k \mid U_1 \ni k, U_1)$.

The conditional probability $\pi_{k|1}$ is defined by the outcome in the first phase. Thus $\pi_k^* = \theta_k \pi_{k|1}$ does not equal the inclusion probability $\pi_k = Pr(s \ni k)$, which is obtained by accounting for all possible U_1, containing unit k, and their probabilities.

An alternative to the weighting estimator in (1) is the "star"-estimator (Särndal et al., 1992)

$$\hat{Y}^* = \sum_s \frac{1}{\pi_k^*} y_k \tag{2}$$

suggested for estimation under two-phase sampling.

The first phase design corresponds to Poisson sampling if the outcomes of x_k are independent among units in the population. Laitila and Olofsson (2011) and Olofsson (2011) treats a two-phase design with Poisson sampling followed by fixed size simple random sampling (SI). They derive the inclusion probabilities

$$\pi_k = \theta_k E(n/N_1 \mid U_1 \ni k) \tag{3}$$

where expectation is with respect to the distribution of N_1, the number of units in U_1.

With proportional sampling $n = f_1 N_1$ a two stage design is obtained with $\pi_k = \theta_k f_1$. Furthermore if $plim_{N\to\infty} N_1/N = \alpha$, a constant, and $n = O(N)$ then $(n/N_1 - E(n/N_1 \mid U_1 \ni k)) \to 0$ in probability as $N \to \infty$. Thus, for larger population sizes N, implying larger N_1, an approximation of inclusion probabilities is

$$\pi_k \approx \theta_k (n/N_1) = \pi_k^* \tag{4}$$

For small N the approximation can not be assumed useful. With cut-off sampling the approximation may work well for estimation of population totals. This may not be so in domains close to the cut-off threshold because of larger variation in domain sizes.

3 Cut-off Sampling in Practice

The idea of leaving a part of the population outside of the sampling frame is to make survey designs more efficient, trading a small bias for reduced estimator variance and/or lower costs. Here the excluded population part is considered less important for estimates of totals. This means that the cutoff is made indirectly with respect to the size of the study variables. Hence, a unit in the sampling frame is not representative of those units not included. However, if a unit is included in U_1 but was excluded from a previous sampling frame, say at time $t-1$, it is representative for the excluded population part at that time. If data for time $t-1$ can be retrieved from the unit, it can be used for estimation of characteristics of the population part excluded at time $t-1$.

Let F_{1t} denote the cut-off sampling frame at time t. Sampling of units from the population U_{t-1} is obtained in two steps:

Step 1 - A QR selection of units in U_{t-1} into F_{1t}.

Step 2 - A sample s_t is drawn from F_{1t} with a probability sampling design.

For simplicity the populations are assumed to stay the same over the time periods t and $t-1$. Deaths of firms may add problems if they belonged to the population part cutoff in time $t-1$. Those firms have zero probability of being sampled in time t. Deaths of firms in $U_{1,t-1}$ poses a problem

if s_t is used for inference on that cut-off population. Births of firms are handled by exclusion with an inclusion indicator.

In the first step units are selected independently with probabilities θ_k. In the second step units are drawn with inclusion probabilities $\pi_{k|1}$. In combination the inclusion probabilities are $\pi_k = \theta_k E_{s_{1t}}(\pi_{k|1} \mid F_{1t} \ni k) = \theta_k \bar{\pi}_{k|1}$, which equals (3) under SI in Step 2.

Let $Z_k = 1(k \in U_{D,t-1})$ be a domain indicator and consider estimation of the domain total $Y_{D,t-1}$. One example of interest is $Z_k = 1(k \in U_{0,t-1})$ and estimation of $Y_{0,t-1}$. Two estimators are the HT estimator

$$\hat{Y}_{D,t-1} = \sum_{s_t} \frac{1}{\pi_k} Z_k y_k \tag{5}$$

and the star-estimator

$$\hat{Y}^*_{D,t-1} = \sum_{s_t} \frac{1}{\pi^*_k} Z_k y_k \tag{6}$$

where y_k is the value of the study variable at time $t-1$ and s_t is the cut-off sample obtained in time t.

Suppose there is an estimate of $\hat{Y}_{1,t-1}$. An estimate of the total Y_{t-1} is obtained by adding an estimate of $Y_{0,t-1}$ using e.g. (6)

$$\tilde{Y}_{t-1} = \hat{Y}_{1,t-1} + \hat{Y}^*_{0,t-1} \tag{7}$$

with $Z_k = 1(k \in U_{0,t-1})$. If the first estimator on the right-hand side is based on the cut-off sample s_{t-1}, the two estimators on the right are independent because the two phase sampling design effectively implies a stratification of U_{t-1} into the strata $U_{1,t-1}$ and $U_{0,t-1}$. It is possible to define several other estimators of $Y_{0,t-1}$ and Y_{t-1} with or without auxiliary variables, e.g. ratio and regression estimators. By setting $Z_k = 1$ the estimators provide estimates of Y_{t-1} directly from the cut-off sample s_t. However, this last estimator will be biased by deaths of firms included in $F_{1,t-1}$.

4 Properties of estimators

The properties of the estimators (5) and (6) are well known and can be found in results 2.8.1 and 9.3.1, respectively, in Särndal et al. (1992). For the HT estimator the second order inclusion probabilities are $\theta_k \theta_l E_{F_{1t}}(\pi_{k,l|1} \mid F_{1t} \supset \{k.l\}) = \theta_k \theta_l \bar{\pi}_{k,l|1}$ $(k \neq l)$.

Using Result 2.8.1 in Särndal et al. (1992) a variance estimator for the HT estimator is

$$\hat{V}(\hat{Y}_{D,t-1}) = \sum \sum_{s_t} \frac{\bar{\pi}_{k,l|1} - \bar{\pi}_{k|1}\bar{\pi}_{l|1}}{\bar{\pi}_{k,l|1}\bar{\pi}_{k|1}\bar{\pi}_{l|1}\theta_k\theta_l} D_k y_k D_l y_l + \sum_{s_t} \frac{1-\theta_k}{\bar{\pi}_{k|1}\theta_k^2} D_k y_k^2 \tag{8}$$

Using Result 3.9.1 (ibid) gives

$$\hat{V}(\hat{Y}^*_{D,t-1}) = \sum \sum_{s_t} \frac{\pi_{k,l|1} - \pi_{k|1}\pi_{l|1}}{\pi_{k,l|1}\pi_{k|1}\pi_{l|1}\theta_k\theta_l} D_k y_k D_l y_l + \sum_{s_t} \frac{1-\theta_k}{\pi_{k|1}\theta_k^2} D_k y_k^2 \tag{9}$$

Both variance estimators takes on the same form and they resemble the variance estimators (3.12) in Lundström and Särndal (1999) and (5) in Chang and Kott (2008) for estimators adjusting for nonresponse. For the HT estimator $\bar{\pi}_{k|1}$ is defined for all units in U_{t-1} and is invariant to F_{1t}. The marginal distribution of s_t equals a joint distribution of two independent trials, $\pi_k = \theta_k \bar{\pi}_{k|1}$. This is an assumption usually made in applications of the QR approach and explains the similarity

between (8) and variance estimators for nonresponse adjusted estimators. The similarity of (9) is due to treating second stage inclusion probabilities conditional on the outcome of the first step, ignoring the invariance.

5 Selection Bias

A critical part in estimation is the assessment of the θ_k, the probabilities of population units being over the cut-off threshold a given year. If they are incorrect validity of estimation results cannot be claimed.

One approach is to utilize observational data on firms moving and not moving over the threshold from one year to another. Table 1 presents data from a fictional Swedish survey on firms below the threshold of 10 employees in 2020 and above nine in 2021. The data were collected from the Swedish business registry, and sole traders - businesses whose organisation number equals the owner's personal number - were excluded. As a study variable register data on turnover is used.

In the left part of the table, the number of firms and mean turnover are presented for all firms (0-9), domains of number of employees $\{0, 1, ..., 9\}$ and for firms with at least one employee (1-9). The right part presents the same variables for the subset of firms moving over the threshold in 2021.

For each size category $\{0, 1, ..., 9\}$, the subset of firms over the threshold in 2021 is here treated as a constant probability Poisson sample from the population units below the threshold in 2020. The sample mean is considered an estimate of the domain population mean in 2020, and associated standard error estimates are given.

For firms with 1 employee a 95% confidence interval $(2287 \pm 1.96 \cdot 355)$ covers the true domain mean 1928. Confidence intervals for the other size classes also covers the true population value, with exception for firms with 2 employees where the true value falls slightly below the lower limit 3728.4. Thus, the separate domain estimates do not contradict an assumption of equal sample probabilities within the size classes.

The mean estimate over all domains (0-9) is more than twice as large to the true value 4249. The standard error is also large yielding the 95% confidence interval 4078 – 15572. Over the domain with at least one employee the mean estimate is close to the true value and its standard error is small.

The difference is explained by the sample obtained of firms with 0 employees. The estimate there is more than 3 times as large than the true value. One explanation here is an erroneous assumption of equal probability among firms of moving over the threshold in 2021. Firms with higher turnover are indicated to have a higher probability. If so the sample is characterized by a selection bias problem.

Selection bias is not indicated by the results for the other size categories. This cannot be excluded however and there may very well be such a problem regarding estimation of other variable totals.

6 Discussion

We have suggested a method of drawing inference on the subpopulation excluded under cut-off sampling when the population is repeatedly surveyed with different probability samples. The idea is to utilize information on population units that are excluded from the cut-off sampling frame in one survey, but turn up in the cut-off sample in another.

Table 1: Data.

Population under CO 2020			Units moved over CO in 2021		
N:o Employees	N:o Firms	Mean Turnover	N:o Firms	Mean Turnover	Standard Error
0	267 401	3 648	413	13 118	5 238
1	92 240	1 928	120	2 287	355
2	42 163	3 572	155	5 757	1 035
3	21 678	5 437	164	6 748	1 712
4	15 386	7 408	214	7 700	1 040
5	11 683	10 423	341	9 201	2 297
6	9 251	11 403	569	10 346	1 439
7	7 565	12 491	842	11 741	1 661
8	6 020	15 162	1 310	15 719	1 663
9	5 019	16 708	2 005	17 488	921
0-9	478 406	4 249	6 133	9 825	2 932
1-9	211 005	5 010	5 720	5 653	360

Samples are obtained via a two-phase design where the first phase is modeled through quasi-randomization describing units moving from being cutoff in one survey to being in the cut-off sampling frame in another survey. The second phase consists of the ordinary sampling design used on the cut-off sampling frame. Given selection probabilities in the first phase, standard theory on two-phase sampling is applicable for inference.

One option in designing the second phase sampling is to account for the purpose of inference on the units cutoff in another survey. One option is to stratify the cut-off sampling frame with respect to exclusion or not from the other frame.

A weakness of the method is the unknown selection probabilities in the first phase. If these are not correctly specified it can introduce selection bias in estimates. Care has thus to be taken when modeling and estimating these probabilities. This is no different from estimating non-response/response probabilities.

However, cut-off sampling is usually applied in business surveys and the event of moving into the cut-off sampling frame stems from a different kind of decision than the one determining a response or non-response. Based on theories on growth of firms (e.g. Penrose 1959) theoretically sound selection models can be developed.

There are several potential uses of the method. As described it can deliver estimates on the set of cutoff units. It can also be used for evaluation purposes where another estimation method is in use. A final suggestion is to use the method combining data from the whole population in estimation of prediction models.

References

R. Benedetti, M. Bee, and G. Espa. A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26(4):651, 2010.

T. Chang and P. S. Kott. Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95(3):555–571, 2008.

H. Elisson and E. Elvers. Cut-off sampling and estimation. In *Proceedings of Statistics Canada Symposium*, 2001.

D. Haziza, G. Chauvet, and J.-C. Deville. Sampling and estimation in the presence of cut-off sampling. *Australian & New Zealand Journal of Statistics*, 52(3):303–319, 2010.

H.-J. Hwang and K.-I. Shin. An improved composite estimator for cut-off sampling. *Communications for Statistical Applications and Methods*, 20(5):367–376, 2013.

J. Knaub. Cutoff sampling - sage. *Encyclopedia of Survey Research Methods*, Vol 1:175–176, 01 2008. doi: 10.4135/9781412963947.n122.

T. Laitila and J. Olofsson. A two-phase sampling scheme and πps designs. *Journal of Statistical Planning and Inference*, 141(5):1646–1654, 2011.

R. J. A. Little. Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54(2):139–157, 1986.

S. Lundström and C.-E. Särndal. Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15(2):305, 1999.

H. L. Oh and F. J. Scheuren. Weighting adjustment for unit nonresponse. In W. G. Madow, I. Olkin, and D. B. Rubin, editors, *Incomplete Data in Sample Surveys*, volume 2, pages 143–184. Academic Press, New York, 1983.

J. Olofsson. Algorithms to find exact inclusion probabilities for 2pπps sampling designs. *Lithuanian Mathematical Journal*, 51(3):425–439, 2011.

E. T. Penrose. *The theory of the growth of the firm*. Basil Blackwell, Oxford, 1959.

C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

prnsamplr: Permanent Random Number Sampling in R

Kira Coder Gylling
Örebro University and Statistics Sweden

Abstract

We present an R package for coordinated survey sampling using permanent random numbers. This tool implements the sampling procedures of the Swedish SAMU system for economic statistics. We present the functionality of the package and show how it can be used to control overlap between samples.

Keywords survey sampling, sample coordination, software package

1 Introduction

1.1 Background

In official statistics production there are two conflicting issues that need to be handled: Response burden and comparability. The former issue requires us to have a low overlap between samples – negative coordination – and the latter issue requires us to have a high overlap – positive coordination. The problem is that the desired coordination cannot be controlled by drawing independent samples for each survey, when the sampling procedures base the samples on independent sets of random numbers.

The Swedish so-called SAMU system for economic statistics, as described by Ohlsson (1992) and Lindblom (2014), solves this problem by associating with each unit in the business registry a so-called permanent random number (PRN) which replaces the random numbers that the sampling procedure generates, and which is supplied to the sampling procedure in order to control the sample overlap between surveys or within surveys over time. A fortunate side effect of this is that the samples are reproducible in case something goes wrong in the sampling process.

1.2 Outline

In Sections 2.1 and 2.2 we describe the two sampling methods that prnsamplr can handle. In Section 2.3 we describe a central mechanism for shifting the PRN's in order to control sample overlap. In Section 3 we give an overview of the included functions and data. In Section 4 we show some results of using the package, and in Section 5 we recapitulate and summarize what we have discussed.

2 Permanent random number sampling

2.1 Stratified simple random sampling

Simple random sampling is arguably the simplest and most easily explained unbiased sampling routine, and is appropriate when every unit in a given stratum should have the same sample inclusion probability. Given sample sizes n_h and a sampling frame with permanent random numbers (PRN's), we only need two steps in the algorithm:

1. Sort the frame ascending along the PRN's within each stratum.
2. The n_h elements with the lowest PRN's are sampled, for each stratum h.

2.2 Stratified Pareto πps sampling

Pareto πps (probability proportional to size) sampling is a more complicated sampling routine than simple random sampling, and is appropriate when big units should have a higher sample inclusion probability than smaller units. A good introduction to Pareto πps sampling is given by Rosén (2000), but a short summary is given below.

Given sample sizes n_h and a sampling frame containing PRN's and a size measure x, the algorithm is shown below.

1. For each item k in the frame, calculate
$$\lambda_k = n_h \frac{x_k}{\sum_{k \in h} x_k}$$
2. If any item has $\lambda_k \geq 1$ then these items get $\lambda_k = 1$, are sampled, and the remaining λ_k in the corresponding stratum h are recalculated as
$$\lambda_k = m_h \frac{x_k}{\sum_{k \in h'} x_k}$$
where $h^{'}$ is the set of elements in h with $\lambda_k < 1$, and m_h is the desired sample size from $h^{'}$, i.e. the number of elements that remain to be sampled after the ones with $\lambda_k \geq 1$ have been marked for sampling.
3. Repeat step 2 until no new $\lambda_k \geq 1$.
4. When no new $\lambda_k \geq 1$, calculate
$$Q_k = \frac{U_k(1-\lambda_k)}{\lambda_k(1-U_k)}$$
for each item k with $\lambda_k < 1$, where U_k are the PRN's.
5. Sort the frame ascending along Q within each stratum $h^{'}$.
6. The m_h elements with the lowest Q are sampled, for each stratum $h^{'}$.
7. Concatenate the partial samples from steps 2 and 6.

It is possible to calculate Q_k for each element, and thus to set it to 0 for elements with $\lambda_k = 1$. A problem that can appear is if more than n_h elements get $\lambda_k = 1$. In this case it is a good idea to revise the stratification plan, and possibly define a take-all stratum.

2.3 Transformation of permanent random numbers

We want to control the overlap between samples, and hence the sample coordination, in order to either spread out the response burden by having a low overlap (negative coordination) or maximize comparability by having a high overlap (positive coordination). The problem is that the sampling algorithms always select the items with the lowest PRN's, which would imply maximum positive coordination at all times, so we might want to e.g. count down from 1 for one sample and up from 0.4 for another.

Lindblom and Teterukovsky (2007) showed that we can transform the PRN's according to a simple modulo-1 operation such that the start point gets moved to zero and the direction turns to up. Then we can utilize the sampling algorithms without modification using the transformed PRN's.

The operations are, for PRN's U and using start point s,

$$U_{new} = (U_{old} - s + 1) \mod 1, \tag{1}$$

$$U_{new} = 1 - ((U_{old} - s + 1) \mod 1) \tag{2}$$

for directions up and down, respectively.

3 Overview of prnsamplr

3.1 Function srs

A straightforward function, implementing the algorithm in Section 2.1 with a handful of lines. Input is the sampling frame together with parameters stratid, nsamp, and prn, which should point to variables on the frame containing stratum information, sample sizes, and the PRN's, respectively. Output is a copy of the frame together with a variable sampled, indicating which units have been sampled.

3.2 Function pps

This is a recursive implementation of the algorithm in Section 2.2: Step 2 is repeated by calling the function on the subset $\{k : \lambda_k < 1\}$. Input is the same as for srs with an extra parameter size, pointing to a variable on the frame that contains the size measure. Output is the same as for srs, with the extra variables lambda and Q.

3.3 Function transformprn

This function implements the modulo calculations in Section 2.3 with two "if" statements. Input is the sampling frame, the name of the variable containing the PRN's, direction for the transformation, and starting point for the transformation. Output is a copy of the sampling frame with the PRN's transformed according to specification, and with the untransformed PRN variable renamed to prn.old.

3.4 Function samp

A wrapper for srs and pps. Input is the sampling frame, a specification of the method to use – srs or pps – and arguments that the functions srs and pps use. Output is the same as for srs or pps.

3.5 Dataset ExampleData

An example dataset with $N = 40,000$ observations in 100 strata with names ranging from st00001 to st00100. Stratum boundaries were drawn from the $\mathcal{U}(1, N)$ distribution, and sample sizes were calculated as $n_h = U \cdot N_h$, where N_h are the stratum sizes and $U \sim \mathcal{U}(0, 1)$. PRN's were drawn from the $\mathcal{U}(0, 1)$ distribution, and a size measure was calculated as $10 \cdot U$ where $U \sim \mathcal{U}(0, 1)$. Stratification information can be found in Appendix A, and code that generated the data can be found accompanying this paper.

4 Usage and examples

In Tables 1 and 2 we show overlaps between six SRS samples and six πps samples, respectively. For the PRN transformations we used the same starting points and directions as in the SAMU system, see Appendix 2 in Ohlsson (1992). Each sample was drawn on the ExampleData dataset that is included in prnsamplr and used the same stratification.

Table 1: Overlap between six SRS samples with varying start point and direction for the PRN transformation.

	U0.0	U0.2	D0.3	D0.7	U0.7	D1.0
U0.0	100%	67.8%	68.9%	72.3%	57.8%	48.5%
U0.2		100%	57.1%	76.7%	48.7%	66.6%
D0.3			100%	50.9%	76.3%	57.6%
D0.7				100%	48.5%	57.8%
U0.7					100%	68.9%
D1.0						100%

Table 2: Overlap between six πps samples with varying start point and direction for the PRN transformation.

	U0.0	U0.2	D0.3	D0.7	U0.7	D1.0
U0.0	100%	78.4%	82.6%	78.4%	72.6%	67.7%
U0.2		100%	75.7%	83.4%	67.7%	75.2%
D0.3			100%	68.9%	80.8%	72.4%
D0.7				100%	67.8%	72.3%
U0.7					100%	83.0%
D1.0						100%

Furthermore we have run the following tests on the data and code:

- Each stratum has only one population/sample size
- $n_h \leq N_h$ in every stratum
- srs: Exactly n_h items are sampled in each stratum
- pps: Exactly n_h items are sampled in each stratum

- pps: λ_k summarizes to n_h in each stratum

The code and data passed each test.

5 Summary and discussion

In this paper we have presented a package for coordinated survey sampling in R. This package allows us to use permanent random numbers (PRN's) to draw coordinated samples using either of two methods: stratified simple random sampling and stratified Pareto πps sampling. Furthermore it allows us to reduce or increase the sample overlap by shifting the permanent random numbers.

We have drawn samples with various transformations of the PRN's and calculated the overlap between them, and we find that the overlap is reduced when the PRN's are counted from opposite directions or from points that are far from each other. This is expected. We also find that the overlap is in general higher for Pareto πps sampling than for simple random samples. This is also expected.

Finally, the functions and included data pass each test we have ran, verifying that the functions are sound and calculate correctly.

References

A. Lindblom. On precision in estimates of change over time where samples are positively coordinated by permanent random numbers. *Journal of Official Statistics (JOS)*, 30(4), 2014.

A. Lindblom and A. Teterukovsky. Coordination of stratified pareto πps samples and stratified simple random samples at statistics sweden. In *Papers presented at the ICES-III, June 18-21, 2007, Montreal, Quebec, Canada*, 2007.

E. Ohlsson. *SAMU : The system for co-ordination of samples from the business register at Statistics Sweden*. Statistics Sweden, Stockholm, 1992.

B. Rosén. *A User's Guide to Pareto πps Sampling*. Statistics Sweden, Stockholm, 2000.

A Tables

Table 3: Sample and population sizes in the example dataset.

stratum	npopul	nsample	stratum	npopul	nsample	stratum	npopul	nsample
st00001	535	324	st00036	81	76	st00071	243	83
st00002	398	261	st00037	155	93	st00072	344	217
st00003	1424	503	st00038	400	225	st00073	991	834
st00004	114	31	st00039	273	144	st00074	369	316
st00005	356	354	st00040	130	129	st00075	141	56
st00006	542	344	st00041	48	25	st00076	300	115
st00007	609	130	st00042	935	639	st00077	5	5
st00008	339	44	st00043	138	84	st00078	58	38
st00009	550	263	st00044	687	165	st00079	161	120
st00010	155	144	st00045	152	40	st00080	257	156
st00011	710	426	st00046	691	504	st00081	195	177
st00012	1330	1299	st00047	35	16	st00082	123	37
st00013	386	283	st00048	16	3	st00083	542	104
st00014	619	221	st00049	37	28	st00084	404	359
st00015	40	18	st00050	141	15	st00085	257	130
st00016	131	20	st00051	459	397	st00086	482	423
st00017	247	4	st00052	166	103	st00087	871	165
st00018	1099	787	st00053	837	467	st00088	125	95
st00019	207	22	st00054	444	146	st00089	214	156
st00020	28	13	st00055	932	423	st00090	225	213
st00021	501	321	st00056	793	397	st00091	38	21
st00022	300	298	st00057	1069	194	st00092	638	455
st00023	68	34	st00058	1181	626	st00093	247	97
st00024	1055	511	st00059	527	40	st00094	393	40
st00025	907	158	st00060	191	54	st00095	187	174
st00026	363	275	st00061	153	33	st00096	873	248
st00027	282	128	st00062	32	10	st00097	399	236
st00028	55	29	st00063	365	327	st00098	637	71
st00029	212	44	st00064	48	22	st00099	1252	1053
st00030	51	12	st00065	259	203	st00100	325	104
st00031	253	151	st00066	742	654			
st00032	1018	586	st00067	228	95			
st00033	317	25	st00068	736	47			
st00034	94	4	st00069	55	19			
st00035	68	44	st00070	205	149			