Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau

by Jean Opsomer

December 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 227-231
Statistics Canada, Catalogue No. 12-001-X

Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau

Jean Opsomer [1]

1. Introduction

The U.S. Census Bureau is one of the largest survey data collection organizations in the world, in addition to its role in the collection of the U.S. Decennial Census data. The two major statistical tools used by the Census Bureau in designing its surveys are stratification and multi-stage sampling. These tools have been successfully implemented starting in the 1940s and have continually been adapted and refined since then.

While this general sampling approach has been very successful, there are increasing concerns about rising survey costs, decreasing response rates and new frame coverage issues (especially related to telephones). At the same time, advances in data collection methods, new data sources and computational tool offer opportunities for considering survey design approaches that would have been unfeasible before. In conjunction with the 2010 Redesign Program currently on-going at the Census Bureau, input was therefore sought from leading academic researchers in innovative sampling methods, as a way to initiate the exploration of possible new approaches to design surveys conducted by the Census Bureau. As a result, Profs. Steve Thompson (Simon Fraser University), Sharon Lohr (Arizona State University) and Yves Tillé (Université de Neufchâtel) were invited to give overview lectures on some of the designs they developed. I was invited to contribute a discussion to each of these lectures.

In the three sections that follow, I will summarize my comments to each of these lectures. My goals in those comments were to highlight the most important aspects of the sampling methods that were presented, to discuss some of the main opportunities for using these designs in the household sampling context, and to identify possible challenges in implementation.

2. Adaptive network and spatial sampling

Prof. Thompson's lecture covered a broad class of designs that includes adaptive cluster sampling, network sampling and adaptive web sampling. Unless I am referring to a specific design within this class, I will refer to these designs as "adaptive sampling" in what follows. A major advantage of adaptive sampling is that it incorporates some of the features of "convenience" sampling approaches such as snowball sampling, including decreased reliance on a sampling frame and the ability to target sampling to portions of the population of particular interest. But unlike convenience sampling, adaptive sampling remains firmly design-based, in the sense of allowing randomization-based finite population estimation and inference.

In adaptive sampling procedures, an initial sample s_0 is drawn according to a probability sampling design $p_0(s_0)$. Based on the characteristics of the elements in s_0 (*e.g.*, presence/absence of features of interest or an enumeration of "links" to other elements in the population), a follow-up sample s_1 is selected from the remaining population, using a conditional sampling design $p_1(s_1 \mid s_0)$. This process is repeated with successive incremental samples $s_2, s_3, \ldots$ until a target criterion such as overall sample size or number of sampling "waves" is reached, and the final sample is the union of each of the successive samples. The specifics on how the waves are drawn varies by adaptive design. Section 2.2 of Thompson's article in this issue and Thompson (2006) provide additional details for adaptive web sampling, a very flexible type of adaptive sampling that includes many of the other designs as special cases.

Because the designs for each of the sampling waves are probability designs, it is possible to obtain valid design-based estimators. A simple estimator for the finite population mean $\mu_N = N^{-1}\sum_U y_i$ is constructed as follows. Based on the initial design p_0 with associated inclusion probabilities π_{0i}, an unbiased estimator for the population mean is given by $\hat{\mu}_0 = N^{-1}\sum_{s_0} y_i / \pi_{0i}$. For each of the subsequent waves $k = 1, \ldots, K$, an unbiased estimator of μ_N is given by $z_k = \sum_{s_{k-1}} y_i + \sum_{s_k} y_i / q_{ki}$, where q_{ki} are conditional inclusion probabilities for wave k (see Thompson (2006) for details on construction of the q_{ki}, and Section 2.4 of Thompson's article in this issue for specific examples). Letting $\hat{\mu}_r = K^{-1}\sum_{k=1}^{K} z_k$, an unbiased estimator for μ_N is obtained as $\hat{\mu} = w\hat{\mu}_0 + (1 - w)\,\hat{\mu}_r$, which is a linear combination of the initial estimator and the mean of the subsequent estimators.

The estimator $\hat{\mu}$ is design unbiased but it depends on the order of the waves in which the sample was obtained. A more precise estimator can be obtained by averaging over all the different orders in which the same sample could have

1. Jean Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877. E-mail: jopsomer@stat.colostate.edu.

been obtained. For small sample sizes, an explicit expression is available for this more efficient estimator, but in general it needs to be approximated by repeated sampling from an appropriately defined Markov chain, and taking the mean of the samples. The exact methods for setting up the chain and drawing the samples are described in Thompson (2006), which also discusses variance estimation for the resulting estimator.

One of the primary advantages of adaptive sampling designs is that they allow the survey organization to focus the sample in portions of interest in the population. This is particularly useful in situations where some of the elements of interest are relatively rare and where they cannot be identified a priori in a sampling frame. Examples of such situations are surveys of hunting and fishing behavior, recent immigrants, home-schoolers, or owners of family-owned businesses. In each of these cases, the elements are quite "diffuse" in the population and no comprehensive frame is generally available. However, it is likely that individuals who are part of this population will be able to provide information on other individuals, so that links can be identified and sampled across different adaptive sampling waves. Note that adaptive sampling can also be used when these types of rare elements are part of a subpopulation of interest within a survey of a larger and non-rare population. For instance, a survey of school children might want to include a stratum of home-schooled children.

Finding relatively rare (sub)populations is a common challenge in surveys, and a number of methods are regularly deployed to deal with this issue. Perhaps the most common sampling design in the context of household surveys is stratified multi-stage sampling. To the extent that relevant PSU-level auxiliary information is available, the survey organization can oversample PSU expected to contain a larger fraction of the groups of interest. An example of such a situation is a survey of African-American males at risk of Parkinson's disease, in which Census tracts with higher African-American population fraction could be oversampled. Another sampling design that can be useful in this context is multi-phase sampling. In this case, the first phase of sampling is used either as a screening sample or as a way to collect relevant auxiliary information, while subsequent phases focus on obtaining the survey data of interest. The Agricultural Resource Management Survey (ARMS) conducted by the USDA follows this design. A sample of all farms is selected in phase 1, in which farm characteristics for the survey year are collected. In later phases, targeted sampled based on the commodities of interest (*e.g.*, dairy, wheat, *etc*) are selected. A third sampling approach that is sometimes useful for obtaining samples of rare (sub)populations is multi-frame sampling. The principle underlying multi-frame sampling is to combine several frames with different coverage characteristics, for instance a "good" frame containing a large proportion of elements of interest but potentially incomplete and a "bad" frame that is comprehensive but contains a low proportion of elements of interest. For instance, a survey of companies in a particular industry might be able to use an industry group membership list as the "good" frame and a general company list as the "bad" frame. For a more in-depth look at multi-frame sampling, see Section 3 below.

Compared to these three designs, adaptive sampling is more flexible and allows finer control over the number and characteristics of elements that are included in the sample, which will often result in improved efficiency and/or lower cost. A drawback of adaptive sampling is that information needs to be collected on the linkages between elements, which can increase respondent burden and data collection cost, and potentially raises confidentiality issues.

Because adaptive sampling frequently relies on "links" between elements in order to define the conditional selection probabilities in the sampling waves, it is also particularly well-suited for surveys that are interested in studying connections between elements in a population. Examples of such situations might be surveys involving transactions or relationships between businesses, surveys of barter/trading behavior of households, and surveys of family network relationships or characteristics.

For a survey organization contemplating adoption of adaptive sampling, a number of issues related to estimation and data dissemination need to be considered. In many cases, the survey data are released in the form of a weighted dataset, and variance estimates are provided in the form of a simplified design description (*e.g.*, strata and PSUs), replicate weights or generalized variance functions. It is also very common for the weights to be calibrated and/or adjusted for non-response. Estimators for adaptive designs are indeed expressible as weighted sample sums, so that a weighted dataset could readily be created even for the Markov chain version of the estimators mentioned above. The choice of how to best provide variance estimates with the dataset is something that still needs to be investigated and might depend on the specifics of the survey. Similarly, how to incorporate calibration and nonresponse adjustments in adaptive sampling estimation is an area where additional research is needed.

3. Sampling with multiple overlapping frames

Prof. Lohr gave a comprehensive overview of general sampling designs and estimation methods when sampling uses multiple frames. Traditional approaches for conducting surveys are increasingly called into question today, because

of increasing costs, decreasing response levels for traditional modes, and increasing concerns for undercoverage of existing sampling frames (*e.g.*, landline telephone numbers reached by RDD). By drawing samples from several frames instead of from a single frame, it is possible to reduce survey costs, improve the coverage of the overall sample, and potentially even increase response rates depending on the specific survey being conducted (for instance, because of improved respondent identifier information in one of the frames).

Multiple frame sampling is a pure randomization-based approach to draw samples, and sampling within the individual frames follows the same methodology as "classical" single-frame sampling. Fully design-based estimation methods for multiple-frame sampling are available, several of which can readily be deployed in the large-scale survey context in which a weighted dataset is the primary output (see below). The key feature of all estimation methods is the estimation of the frame overlap, which is typically unknown but needs to be accounted for. This is done by, for each frame, constructing design-based estimators for the subpopulation(s) of elements that also fall in the other frame(s). The estimators for the characteristics of the frame intersection(s) then need to be combined across frames. Existing methods differ in how they combine these estimators, with the simplest methods using sample-size weighted averages and more complex estimators weighting by estimates of the precision of the individual estimators.

Sampling from multiple frames is particularly applicable in cases where no single frame is available that covers the whole population. Typical examples of such situations are RDD sampling, where an increasing fraction of the population is not reachable through a landline telephone number, surveys of professionals or businesses with partial listings available from vendors or professional organizations. Other situations in which multiple frame sampling might be applicable are surveys of rare subpopulations that exist within a larger population. An overall frame for the population exists, but screening respondents for whether they belong the the subpopulation is time-consuming and expensive. An alternate frame containing a much higher proportion of elements from the subpopulation of interest is sometimes available, but if the coverage of that frame is incomplete, the survey organization might not be willing to rely on it for fear of not obtaining a valid sample. Combining the alternate but incomplete subpopulation frame with the complete but inefficient population frame might be both cost-effective and statistically defensible. Examples of surveys of such subpopulations are surveys of hunting and fishing, where a license frame often exists but it might be incomplete or out of date. This multiple frame approach might also be useful for a survey of the general population, as a way to increase the sample size within certain subpopulations of particular interest. For instance, in a general survey of farms, it might be of interest to produce estimates for organic farms, which only represent a small fraction of farms but with many of those listed in organic business directories. Section 1 of Lohr's article in this issue gives several additional examples of the wide applicability of multiple frame surveys.

As noted above, estimation methods involve the construction of estimators for the frame intersection subpopulation, which requires selection of a weighting method for the estimators obtained from the different frames. Weighting methods that rely on estimating the precision of these estimators might be preferred from an efficiency perspective. However, they are somewhat problematic to implement in practice, because the resulting weights can vary for different variables in the survey. More practical approaches will forego some efficiency in order to be able to have single weights for all survey variables, a key feature emphasized repeatedly in Lohr's article in this issue. The *pseudo-maximum likelihood* (PML) method of Skinner and Rao (1996) produces a single set of weights and is recommended by Lohr as the method of choice for single surveys, while a simpler fixed-weight approach is preferable for longitudinal surveys.

While the basic methodology for constructing design-based estimators for multiple frame sampling is in place today, there is still a need for further research in approaches for applying calibration and nonresponse adjustment in this context. Because it is possible to apply those adjustments at the individual frame level, the population level, or both levels (depending on the available auxiliary information), an investigation of the properties of the estimators under these different scenarios would be very useful, and should be used to develop guidelines for survey practitioners. Section 3 of Lohr's article in this issue discusses some initial results in this area.

Variance estimation methods for multiple-frame estimators have been developed and are reviewed in Section 4.2 of Lohh's article, and include both linearization and replication approaches. An important practical issue in the use of the linearization approach is that it requires access to the frame identification for all the elements in the sample, because it involves separate estimation of the variance in each frame. This might be undesirable for the survey organization producing the data, for reasons of data confidentiality. In the case of replication methods such as jackknife and bootstrap, it is possible for the survey organization to create sets of replicate weights that do not require disclosure of the frame identity of individual sample elements to the data users. Lohr (2007) recommends the *combined bootstrap* approach for inference for multiple frame sampling.

As an alternative, the *grouped jackknife* of Kott (2001) could also be considered.

Implementing multiple frame sampling surveys can be more challenging than single-frame surveys. There needs to be awareness for the increased potential for nonsampling errors, as discussed in Section 5 of Lohr's article, especially if the data collection modes or protocols vary across frames. For instance, sampled elements in one frame get an advance letter, while those in another frame receive a "cold call" because of lack of address information. It is also possible that the nonresponse characteristics differ across frame, so that separate adjustments are required. Finally, in many cases the elements present in the different frames might have different characteristics (*e.g.*, organic farms belonging to a national organic business association *vs.* those that do not). In all those cases, attention to frame-specific effects and careful weight construction are required in order to obtain valid survey estimators. On the other hand, the presence of multiple frames provides opportunities for measuring nonsampling errors, because they entail multiple samples from the same population. For instance, it might be useful to perform "cold calls" for a portion of the selected elements in the frame with addresses to evaluate mode effects.

4. Balanced sampling with the cube method

The presentation by Prof. Tillé covered the fundamentals of balanced sampling and described the *cube method*, which he developed as a practical algorithm implementing the drawing of balanced samples. The goals of balanced sampling designs are to maintain the representation of the population structure in the sample (hence the term "balance"), and to improve the efficiency of survey estimators. Today, most survey statisticians apply stratification as the primary tool to achieve these two goals. Stratification achieves balance by forcing the sample composition to match the stratum allocation, and improves the efficiency of estimators by removing the component of variance due to between-stratum differences. Systematic sampling is also used to achieve these goals, most commonly in natural resource surveys. In this case, the sample composition matches the population composition exactly along the sorting variable, and approximately for any variable correlated with the sorting variable. Efficiency is gained because sample moments of the variables of interest (approximately) match population moments. While both approaches are widely used and work well, they are relatively inflexible. Stratification often involves dividing the population into "cells" defined by the intersection of stratification variables, which might lead to a proliferation of many small cells with corresponding small sample sizes. Systematic sampling is a highly constrained form of sampling with limited amount of flexibility in sample construction, and with the additional issue of the lack of a design-based variance estimator.

Balanced sampling can be viewed as a generalization of stratification. Under this interpretation, stratified samples are drawn with given probabilities of inclusion for all the population elements, but subject to constraints on the sample size in each stratum. In balanced sampling, the stratification constraints are replaced by constraints of the form $\sum_s \boldsymbol{x}_i / \pi_i = \sum_U \boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is a vector of *balancing variables*. When the $\boldsymbol{x}_i$ are stratum indicators, balanced sampling is the same as stratification, but any categorical or continuous variables (or combination thereof) can be used, which provides a high degree of flexibility in sample construction.

As noted above, the cube method is an algorithm that draws balanced samples given a set of inclusion probabilities and constraints. If exactly balanced samples exist in the population, the algorithm will try to select one of them. If no sample can be found that has the postulated inclusion probabilities and satisfies the balancing constraints exactly, it will attempt to come as close as possible to satisfying the constraints. The cube method requires that the balancing variables $\boldsymbol{x}_i$ be known for all elements in the population. Depending on the survey context, this requirement might represent a key limitation on the applicability of balanced sampling.

Despite the fact that balancing on population-level auxiliary variables is done at the design stage, it seems likely that in practice, calibration and other weight adjustments such as for nonresponse will still often be required. In fact, Tillé recommends the combination of balancing and calibration as the most efficient strategy (see Section 7.4 of Tillé's article in this issue). The theoretical properties of estimators that are both balanced and calibrated still needs to be fully worked out, however.

While balanced sampling maintains the inclusion probabilities of the elements in the population, it is clear that the presence of the balancing constraints affects their *joint* inclusion probabilities and hence the variance of the estimators. This topic is addressed in Section 6 of Tillé's article. Deville and Tillé (2005) showed that, under certain conditions, the variance of balanced sampling estimators can be approximated by a linearization-type variance, which depends on the residuals of a linear regression of the survey variables on the balancing variables. While this is an important and useful result, it does not lead to a variance estimation approach that is applicable to all survey applications. One issue is that variance estimation based on this method requires access to the balancing variables for all the survey respondents, and these might not be made

publicly available as part of the survey dataset. In this context, a replication-based method might be particularly attractive, because it would not require releasing these variables. However, no such method is currently available.

Balanced sampling has close connections with *rejective sampling*, which aims to achieve the same goals. In rejective sampling, a sample is drawn with prespecified inclusion probabilities and the sample is accepted or rejected based on whether it is within a given tolerance level of a balancing constraint. If the sample is rejected, the procedure is repeated until a sample is found that falls within the tolerance level. While rejective sampling has a long history, Fuller (2009) described some asymptotic theory that showed that asymptotically, his version of rejective sampling was approximately equivalent to balanced sampling.

5. Closing remarks

The methods covered in the three lectures are remarkably complementary. Adaptive designs make it possible to obtain randomization-based, statistically valid samples for populations that have traditionally been difficult to sample efficiently. Very little frame information is required to draw such a sample, but a significant amount of effort has to be expended during the data collection in order to identify and follow the "links" among the elements, and draw the successive samples. In contrast, balanced sampling is useful when very detailed frame information is available, and in that situation, it allows for highly customized and efficient sample designs. Once a balanced sample is drawn, the data collection can proceed in the same manner as for traditional surveys. Multiple frame sampling covers an intermediate case, in the sense that no single good frame exists but several partial frames are used to "offset" each other's weaknesses. Separate samples are drawn from each frame, and data collection proceeds as usual, except for that fact that it is necessary to determine which frame(s) each sampled respondent belong to.

Combined with the existing approaches already in use, these three new sampling methods have the potential to greatly increase the flexibility with which samples can be customized for specific applications, to reduce survey costs and to increase the precision of survey estimators.

References

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 2, 569-591.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Lohr, S. (2007). Recent developments in multiple frame surveys. In *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, 3257-3264.

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Thompson, S.K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224-1234.