Proceedings of Statistics Canada Symposium 2014
Beyond traditional survey taking: adapting to a changing world

Explorations in Non-Probability Sampling Using the Web

J. Michael Brick[1]

Abstract

Although estimating finite populations characteristics from probability samples has been very successful for large samples, inferences from non-probability samples may also be possible. Non-probability samples have been criticized due to self-selection bias and the lack of methods for estimating the precision of the estimates. The wide spread access to the Web and the ability to do very inexpensive data collection on the Web has reinvigorated interest in this topic. We review of non-probability sampling strategies and summarize some of the key issues. We then propose conditions under which non-probability sampling may be a reasonable approach. We conclude with ideas for future research.

Key Words: Inference, representativeness, self-selection bias

1. Introduction

Probability sampling is generally accepted as the most appropriate method for making inference that can be generalized to a finite population. This method has a rich history and a solid theoretical foundation that has been proven to be effective in numerous empirical studies. With a probability sample, every unit in the population has a known, non-zero chance of being sampled, and in the design-based framework these probabilities are the basis for the inferences (Hansen, Hurwitz, and Madow, 1953; Särndal, Swensson, and Wretman, 1992; Lohr, 2009). Almost all official statistics use this methodology and many national statistical offices require probability sampling for making inferences.

But probability sampling is not the only method for drawing samples and making inferences. In fact, during the 20th century the shift to probability sampling began well after the publication of the theoretical basis for probability sampling by Neyman (1934). Quota samples that only require samples meet target numbers of individuals with specific characteristics such as age and sex have been used for many years, especially in market research. Stephan and McCarthy (1958) review this method of non-probability sampling in election and other types of surveys in the middle of the 20th century in the U.S.

The type of nonprobability sampling used in commercial and market research practice changed dramatically in the last twenty years as access to the Internet became more common in North America and many parts of Europe. Especially in the last decade, online surveys – with respondents drawn from "opt-in" panels – have become extremely popular. The vast majority of these surveys are not probability samples. The reason for their popularity is the low cost per completed interview, with costs much lower than even low-cost probability sample survey methods such as mail. Some of the attractiveness of probability samples has also been lost due to rising nonresponse (Brick and Williams, 2013) and concerns about the frame undercoverage. These issues raise concerns about the validity of inferences from a probability sample. Even staunch advocates of probability sampling have been forced to confront the issue of whether a probability sample with a low response or coverage rate retains the highly valued properties of a valid probability sample (Groves, 2006).

The next section summarizes some important findings from a non-probability sampling task force commissioned by the American Association of Public Opinion Research (AAPOR). This serves as a prelude to some current methods and avenues for further research.

[1] Westat and JPSM, 1600 Research Blvd. Rockville, MD USA 20850

2. Task Force Report

The AAPOR Task Force was asked "to examine the conditions under which various survey designs that do not use probability samples might still be useful for making inferences to a larger population." The task force report, completed in early 2013, can be downloaded from that organization's web site (www.aapor.org). Baker et al. (2013) summarized the report; comments from five experts in the field and a rejoinder are published in the same issue of the journal. Rather than repeat the findings again, we have chosen a few critical ones (in quotes below) that have been the topic of several discussions subsequent to the publication of the report and its summary.

"Unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling." The point of this statement is sometimes misunderstood. The intent is to highlight that talking about all non-probability methods together is of little value because the methods are so different. Issues and concerns about respondent driven sampling methods and opt-in Web panels are very different. Even within the generic term of opt-in Web panels the methods used to select respondents and produce estimates may be distinctive.

"The most promising non-probability methods for surveys are those that are based on models that attempt to deal with challenges to inference in both the sampling and estimation stages." This finding is more hypothesized than based on empirical results. In many ways it parallels the expectation that responsive design may lead to lower nonresponse bias in probability samples (Lundquist and Särndal, 2013). The rationale is that a more diverse set of respondents will reduce biases, given the equivalent weighting scheme. While this seems reasonable, it has not yet been consistently validated in either probability samples (with responsive design) or non-probability samples.

"If non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality." No one study or set of studies can prove that a data collection and estimation strategy will produce estimates that are reasonable for most uses. For example, the *Literary Digest* had correctly predicted the winner in every election from 1920 until its infamous error in predicting Landon as a landslide winner in 1936. Empirical results are important but there must be a set of principles that support the data collection and estimation process so that failures can be explained. Probability sampling has such a foundation, and the theory is why when probability sample estimates are not accurate the failures can be link to deviations such as nonresponse and the theory does not have to be discarded.

"Non-probability samples may be appropriate for making statistical inferences, but the validity of the inferences rests on the appropriateness of the assumptions underlying the model and how deviations from those assumptions affect the specific estimates." The members of the task force believed this finding would be the most controversial (Bethlehem and Cooben, 2013). While this was a contentious issue when the report was first released, we found many agreed with the position, including most of the experts in the discussion of the journal article.

Another area of statistical research that is in much the same position as non-probability sampling is observational studies. Madigan et al. (2014) commented that

"Threats to the validity of observational studies on the effects of interventions raise questions about the appropriate role of such studies in decision making. Nonetheless, scholarly journals in fields such as medicine, education, and the social sciences feature many such studies, often with limited exploration of these threats, and the lay press is rife with news stories based on these studies…the introspective and ad hoc nature of the design of these analyses appears to elude any meaningful objective assessment of their performance..."

Despite these concerns about of validity observational studies, researchers in that area understand the critical importance of the role of these studies and are focused on assessing what can be done to improve the science. Our view is that the same sense of urgency to improve non-probability samples is needed, rather than simply disregarding all forms of non-probability sampling as unsound.

There is evidence that work in inference from non-probability samples is continuing, although much of it is more empirical than theoretical. For example, Barratt, Ferris and Lenton (2014) use an online sample to estimate the size and characteristics of a rare subpopulation. Their evaluation method is similar to many previous studies; they compare the online sample estimates to those of a probability sample and find some important differences. Even though there are differences, they suggest the online sample can be useful when combined with a probability sample.

Wang et al. (2014) use a sample of Xbox users that is clearly not representative of voters in the U.S. elections and use model-based estimation methods to produce election predictions. They show the estimates have small biases despite the problems with the sample. These types of applications and investigations are extremely valuable, even though no single study may provide the theoretical foundations we believe is essential. It is possible that the examinations of many such applications may provide the fuel that sparks a new ways of thinking about foundational issues.

3. Fitness and Conditions for Use

Another point raised by the Task Force was that the usefulness of estimates from non-probability samples (or any sample for that matter) has to be tied directly to the intend purpose – called "fit for use." Some ways of evaluating whether a non-probability sample should even be considered at this time are described below and they are tied to this concept of fitness. We suspect these conditions will change as we gather more information about specific methods of non-probability sampling and estimation.

The preponderance of empirical results has shown that probability samples generally have lower biases than non-probability samples (Callegaro et al., 2014). However, there are situations in which non-probability samples may be the better choice. Brick (2014) suggested three criteria to consider for using non-probability instead of a probability sample. The three conditions are:

a. The cost of data collection for the non-probability should be substantially lower than the cost for the probability sample.
b. The estimates should not have to be extremely accurate to satisfy the survey requirements.
c. When the target population is stable and well-understood non-probability sample be considered even when higher levels of accuracy are needed.

Condition (a) is necessary, but not sufficient. In other words, there are situations where a low-cost non-probability sample may be worse than no information at all because it results in actions that are counter-productive. Hansen, Madow, and Tepping (1983) argued that the cost differential for a model-based sample (a non-probability sample) was not much lower than for a probability sample, so there was little reason to not do a probability sample. The Internet has changed the cost structure dramatically since 1983, and now the costs of a non-probability sample can be very much lower than for a probability sample.

Conditions (b) and (c) are directly related to fitness for use. If estimates that approximate the population quantity are all that is needed, then a low-cost non-probability sample may be appropriate. Even in the comparisons that found non-probability sample estimates were not as accurate as those from probability samples, the non-probability samples estimates were similar to those from the probability samples across a broad range of online sampling strategies. Condition (c) acknowledges that some non-probability samples have been consistently accurate, but those are where there is a stable population and powerful auxiliary data exist. Some establishment surveys may fit into this category because of their stability and the presence of important auxiliary variables available on the sampling frame (Knaub, 2007). Election studies in the U.S. also fall into this realm, especially because there are well-known and powerful predictors of election behavior. It is worth noting that the election outcome is a single outcome, and estimates for other characteristics from these non-probability samples have not been closely evaluated. Hansen (1987) gives a cautionary note that shows that stability cannot be assumed when society or the target population are undergoing changes.

One of the features that sets probability sample empirical results in stark contrast to those from non-probability samples in general is its ability to produce a wide array of estimates with small or reasonable biases. It is this multipurpose capacity that is most lacking from non-probability samples. One reason for this is associated with the modeling activity required in non-probability samples. For example, it is practical to carefully model a particular outcome (in election studies this would include multiple election contests that have similar relationships between predictors and outcome) and to do so with more precision and possibly less bias than with a standard probability sample. The same type of modeling effort is used in small area estimation models where the sample size is too small, and may even be zero, to produce reliable estimates using design-based methods. We are not aware of small

area models that have been proposed for a wide range of statistics and purposes. This is a significant challenge for non-probability samples using Web samples.

These issues that are critical of non-probability samples do not imply that probability sampling is without serious problems of its own. Probability sampling assumptions fail to hold in practice and nonresponse and coverage errors are at the heart of the failure. Measurement errors, often an even greater source of errors, affect both probability and non-probability samples. For example, the empirical estimates of the differences from population totals shown in Messer and Dillman (2011) clearly show that a probability sample alone does not make survey estimates immune from large biases.

4. Non-probability Online Sampling Methods

The Task Force discusses and defines a variety of online sampling methods used in non-probability samples. Callegaro et al. (2014) covers these in more detail, so the full detail is not discussed here. Many opt-in surveys use "river sampling" and "router" sampling methods and these capture respondents from various Web pages and send them to a survey. As a result, these and similar methods are subject to selection biases that are very complex and attempts to compensate for the biases, such as using propensity weighting adjustments, are suspect.

Web panels that use respondents from these types of sampling methods are subject to biases from exactly the same sources. The panels do have some additional beneficial features. For example, data from the "profile" of the panel members may be used in adjusting the estimates and this may be useful in dealing with panel nonresponse. This could lead to smaller biases in estimates of change. These panel "profiles" may also be helpful in reducing the original selection biases, but this is largely unproven. Selection biases are very difficult to understand and deal with effectively.

Following the work in observational and epidemiological studies, some opt-in surveys have begun to use matching (Rivers and Bailey 2009). As noted previously, the AAPOR Task Force viewed this approach has having significant promise, but the less than stellar record of observational studies must be taken into account. The ability to use matching in combinations with other online sampling and weighting methods is a challenge because of the need to estimate so many characteristics and relationships in surveys. In observational studies, there are generally a much smaller number of key outcome statistics and the feasibility of careful modeling of each outcome is greatly enhanced.

Another approach to non-probability sampling methods that has been explored is combining a large online non-probability sample with a small probability sample (often called a reference sample) to adjust for biases in the non-probability sample. The major obstacle with this method is that the effective sample size is a function of the probability sample, so the large sample of the non-probability sample is essentially shrunken to the size of the probability sample. These issues are discussed by several researches, going back at least to Cochran (1977) and more recently Bethlehem (2008). Dever, Rafferty, and Valliant (2008) show that this approach can reduce biases, but reinforce the loss of precision associated with the reference sample. A related approach is a hybrid sampling combination of probability and non-probability samples (Berzofsky, Williams, and Biemer, 2009), but this method has not yet garnered much interest.

5. Path Forward

The future of online sampling is not clear and the multiple approaches that have been attempted in the past decade are a testament to its evolving nature. During this same time, there has also been a search in probability sampling to deal with its challenges. One possible path to the future is trying to leverage all of these efforts to improve the empirical performance and theoretical basis for both methodologies.

For example, research in sampling and data collection such as balanced sampling, responsive design, adaptive design, R-indicators and other measures of representativeness could be an avenue toward better methods for both

types of studies. Coverage adjustment, estimated control totals, and nonignorable nonresponse adjustments might be estimation methods of practical importance.

In addition to these survey methods, other areas offer ideas that should be considered. Causal inference has a rich tradition of dealing with selection bias and newer methods are continuing to be introduced and explored. Fields such as cognitive psychology and behavioral research have also expanded from when they were first introduced as a toolkit into survey research in the 1970s. Of course, information science has undergone a revolution and new areas, such as Big Data, and old ideas with new technologies, like administrative records, could provide new insights and need to be considered.

While this is not a recipe for improving non-probability sampling inference, it does imply that research is possible and essential. Tools and methods exist that may help provide the framework for making inferences from non-probability samples, but without innovative research we will remain in the current muddle.

References

Baker, Reg, J. Michael Brick, Nancy A. Bates, Mike Battaglia, Mick P. Couper, Jill A. Dever, Krista J. Gile, and Roger Tourangeau. (2013), "Summary Report of the AAPOR Task Force on Non-probability Sampling", *Journal of Survey Statistics and Methodology* 1, 90-143.

Barratt, Monica J., Jason A. Ferris, and Simon Lenton. (2014), "Hidden Populations, Online Purposive Sampling, and External Validity Taking off the Blindfold", *Field Methods*, 1525822X14526838.

Berzofsky, Marcus, Rick Williams, and Paul Biemer. (2009), "Combining Probability and Non-Probability Sampling Methods: Model-Aided Sampling and the O*NET Data Collection Program", *Survey Practice*, 2.6, Downloaded December 15, 2015 from www.surveypractice.org/index.php/SurveyPractice/article/view/184/html.

Bethlehem, Jelke. (2008), "Can we make official statistics with self-selection web surveys?", In *Proceedings of Statistics Canada Symposium*.

Bethlehem, Jelke, and Fannie Cooben (2013), "Web Panels for Official Statistics?" *Proceedings 59th ISI World Statistics Congress*, 25-30 August 2013, Hong Kong. Downloaded on May 1, 2014 from http://2013.isiproceedings.org/Files/IPS064-P1-S.pdf.

Brick, J. M. (2014), "On Making Inferences from Non-Probability Samples", Washington Statistical Society 2014 President's Invited Seminar, Washington DC (March 26, 2014).

Brick, J.M., and Douglas Williams. (2013), "Explaining rising nonresponse rates in cross-sectional surveys", *ANNALS of the American Academy of Political and Social Science*, 645, 36-59.

Callegaro, Mario, R. Baker, J. Bethlehem, A. Göritz, J. Krosnick, and P. Lavrakas, eds. 2014. *Online Panel Research: A Data Quality Perspective*. John Wiley & Sons.

Cochran, William. (1977), *Sampling techniques*. New York, Wiley and Sons.

Dever, Jill A., Ann Rafferty, and Richard Valliant. (2008), "Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?", *Survey Research Methods*, 2, 47–62.

Groves, Robert M. (2006), "Nonresponse rates and nonresponse bias in household surveys", *Public Opinion Quarterly*, 70, 646-675.

Hansen, Morris H. (1987), "Some History and Reminiscences on Survey Sampling", *Statistical Science*, 2, 2, 180-190.

Hansen, Morris H., William N. Hurwitz, and William G. Madow. (1953), *Sampling survey methods and theory*. *Vol I*, John Wiley and Son Inc., New York.

Hansen, Morris H., William G. Madow, and Benjamin J. Tepping. (1983), "An evaluation of model-dependent and probability-sampling inferences in sample surveys", *Journal of the American Statistical Association* 78, 384, 776-793.

Knaub J. (2007), "Cutoff Sampling and Inference", *InterStat*, April.

Lohr, Sharon. (2009), *Sampling: design and analysis*. Cengage Learning.

Lundquist, Peter, and Carl-Erik Särndal. (2013), "Aspects of responsive design with applications to the Swedish Living Conditions Survey", *Journal of Official Statistics* 29, 557-582.

Madigan, David, P. Stang, J. Berlin, M. SchuemieJ, M. Overhage, M. Suchar, B. Dumouchel, A.Hartzema, P. Ryan (2014), **"**A Systematic Statistical Approach to Evaluating Evidence from Observational Studies"**,** *Annual Review of Statistics and Its Application*, 1, 11 -39

Messer, Benjamin L., and Don A. Dillman. (2011), "Surveying the general public over the Internet using address-based sampling and mail contact procedures", *Public Opinion Quarterly* 75, 429-457.

Neyman, Jerzy. (1934), "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, 558-625.

Rivers, Douglas, and Delia Bailey. (2009), "Inference from Matched Samples in the 2008 U.S. National Elections", Paper presented at the 64th Annual Conference of the American Association for Public Opinion Research, Hollywood, Florida, May.

Särndal, Carl-Erik, Bengt Swensson, and Jan Wretman. (1992), *Model assisted survey sampling*. Springer.

Squire, Peverill. (1988.), "Why the 1936 Literary Digest Poll Failed?", *Public Opinion Quarterly* 52, 125–33.

Stephan, Fredrick. F., and Philip J. McCarthy. (1958), *Sampling opinions: An analysis of survey procedure*, John Wiley and Son Inc., New York.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. (2014), "Forecasting Elections with Non-Representative Polls", *International Journal of Forecasting*), doi:10.1016/j.ijforecast.2014.06.001.