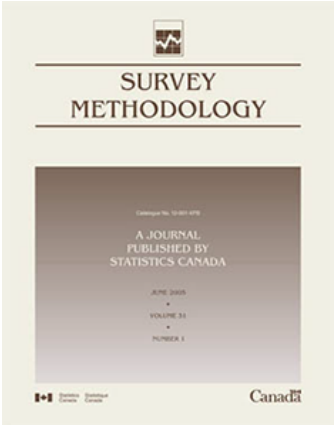
Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Statistical inference with non-probability survey samples”



by Sharon L. Lohr

Release date: December 15, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 331-338
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical inference with non-probability survey samples"

Sharon L. Lohr[1]

Abstract

Strong assumptions are required to make inferences about a finite population from a nonprobability sample. Statistics from a nonprobability sample should be accompanied by evidence that the assumptions are met and that point estimates and confidence intervals are fit for use. I describe some diagnostics that can be used to assess the model assumptions, and discuss issues to consider when deciding whether to use data from a nonprobability sample.

Key Words: Convenience sample; Diagnostics; Imputation; Probability sample; Survey quality; Survey weights.

1. Introduction

Many thanks to Changbao Wu for his stimulating review and assessment of methods for making inferences from nonprobability samples. I especially appreciate his thoughtful examination of the strong assumptions needed to derive the bias and variance of estimates.

Wu reviews three approaches for estimating the finite population mean μ_y of a variable y that is measured in a nonprobability sample S_A of size n_A. Because this sample is not representative of the population (and hence the sample mean $\bar{y}_A$ is likely biased for estimating μ_y), each approach relies on information from a high-quality probability sample S_B of size n_B: S_B does not measure y but it contains a set of auxiliary variables $\mathbf{x}$ that are also observed in S_A.

In the model-based predictive approach, a model is developed on S_A to predict y from $\mathbf{x}$. The mass imputation (MI) estimator, for example, uses the model to impute an estimate y_i^* of y_i for every member of the probability sample S_B. Then the population total of y is estimated by $\sum_{i \in S_B} d_i^B y_i^*$ where d_i^B is the design weight of unit i in S_B.

In the inverse propensity weighting (IPW) approach, a model is developed predicting the probability π_i^A that population unit i appears in S_A as a function of $\mathbf{x}$. Then unit i in S_A is assigned weight $w_i^A = 1 / \hat{\pi}_i^A$ and the population total is estimated by $\sum_{i \in S_A} w_i^A y_i$.

Wu also reviews a "doubly robust" estimator of μ_y that, by combining the predictive and IPW estimators, is approximately unbiased under the assumptions if either model is correctly specified. In this discussion, I will concentrate on the predictive and IPW approaches because these methods generalize more easily for multivariate analyses and estimating population characteristics other than means.

In Section 2, I explore assumptions needed for inference from nonprobability samples and diagnostics for assessing them. Then, in Section 3, I look at some questions to ask when deciding which approach (if any) to use for inference.

1. Sharon Lohr is Professor Emerita of Statistics at Arizona State University. E-mail: sharon.lohr@asu.edu.

2. Model assumptions and diagnostics

Probability sampling gained widespread use after the theory was developed in the 1930s and 1940s because it provided a mathematically justified solution to the problem of how to generalize from a sample to a population. Under minimal assumptions, a full-response probability sample produces approximately unbiased estimates of population quantities, accompanied by confidence intervals that have approximately correct coverage probabilities. It is the *only* method that is guaranteed to produce accurate confidence intervals without making assumptions about the unsampled members of the population. A probability sample is representative because of the procedure by which it is drawn.

All other methods require huge assumptions. The major assumptions for the predictive and IPW methods, given in Section 2.1 of Wu's article, are: (A1) y and the random variable indicating participation in S_A are independent given $\mathbf{x}$, (A2) every unit in the population has $\pi_i^A > 0$, and (A3) the random variables indicating participation in S_A are independent given $\mathbf{x}$. These assumptions imply that the auxiliary information $\mathbf{x}$ is rich enough to develop inverse propensity weights that remove selection bias for y, and that a model developed on S_A to predict y from $\mathbf{x}$ will also apply to units not in S_A.

Statistical properties of the estimators are developed assuming that (A1) - (A3) are true and that the models adopted for weighting or imputation are correctly specified. Under those conditions, the estimated population mean is approximately unbiased with variance given by the appropriate theorem. But, as Wu points out, that variance estimate is conditional on the assumptions being satisfied; if the assumptions are not met, it will severely underestimate the true mean squared error and give a misleading impression of the estimate's trustworthiness. If n_A and n_B are large but (A1) is violated, the bias might be 10 percentage points but the reported standard error of an MI or IPW estimate will be close to zero. In practice, many nonprobability samples will violate the assumptions: Mercer, Lau and Kennedy (2018) found, when weighting online opt-in samples with rich auxiliary information, that "even the most effective adjustment strategy was only able to remove about 30% of the original bias".

The assumptions cannot be fully tested because they involve missing data – population members missing from S_A and y values missing from S_B. But, as with nonresponse adjustments in probability samples (Lohr, 2022, Chapter 8), one can perform model checks and diagnostics using available information, with the recognition that these might not catch all model deficiencies.

Compare statistics from the nonprobability sample with those from other data sources

Wu suggests comparing empirical distribution functions of variables in $\mathbf{x}$ from S_A with the survey-weighted empirical distribution functions from S_B. Differences may indicate that observations in S_A have unequal propensity scores or that the $\mathbf{x}$ variables are measured differently in S_A than in S_B (see Section 3). One can also compare empirical distributions from S_A with those from another probability survey S_C.

If IPW is used, one can also compare propensity-score-weighted empirical distribution functions from S_A with those from S_B and other surveys. This should be done only for variables not used in the

weighting, since the propensity score weights have already adjusted for imbalances in weighting variables. Dutwin and Buskirk (2017), for example, constructed propensity weights for a nonprobability sample through raking on marginal totals and then compared the cross-tabulations of those raking variables.

Wu also suggests treating a variable z that is measured in both S_A and S_B as a response variable, and comparing conditional models for $z \mid \mathbf{u}$ fitted on S_A and S_B, where $\mathbf{u}$ is a subset of $\mathbf{x}$ (excluding z). Differences in the two models can indicate that z is needed as an auxiliary variable, and may also raise questions of how well the set of measured auxiliary variables satisfy assumption (A1).

In an example from Kim, Park, Chen and Wu (2021), the estimated percentage of persons who volunteer was 24.8% from the Current Population Survey (the gold-standard estimate), but the MI and IPW estimates from S_A were both close to 50% with reported standard error less than one percentage point. The standard error, computed under the model assumptions, did not account for the selection bias of S_A with respect to volunteerism – a bias that could not be removed using demographics, home ownership, and medical insurance as model covariates.

Compare results from the IPW and MI approaches

An alternative to using the doubly robust estimator for analysis is to use each model to identify potential deficiencies of the other. Possible investigations include comparing the empirical distribution of y from S_A (using the inverse propensity weights) with the empirical distribution of y^* from S_B (using the imputed values and the survey weights). Similarly, as suggested by Chipperfield, Chessman and Lim (2012), one can compare estimated domain means from S_A and S_B for a set of domains $d = 1, \ldots, D$. One might also compare imputations for y fit to the unweighted data set S_A with imputations developed on S_A with inverse propensity weights.

Simulation studies are valuable for checking the small-sample behavior when the assumptions are met, but are of limited value for exploring sensitivity to model assumptions. These explore model deviations devised by the investigators, but real surveys can diverge from the model in many unanticipated ways.

Perform model diagnostics

Of course, for either the IPW or model-predictive approach, analysts should employ standard regression diagnostics such as examining residuals and influential observations to examine model fit and sensitivity to outliers, and document the checks that were done.

For the IPW approach, it is also desirable to examine characteristics of the final weights. The coefficient of variation of the weights provides a rough measure of the amount of adjustments that were needed to make sample S_A "representative". A low coefficient of variation, however, does not necessarily mean the sample is representative; this may merely reflect inadequacy of the available auxiliary information for developing weights. For example, suppose a quota sample from an opt-in internet panel is drawn to match the population with respect to the auxiliary variables. The inverse propensity weights will have little variation because the $\mathbf{x}$ variables were used to form the quota classes, but the sample may still produce biased estimates of y variables such as internet usage or volunteering.

The graphical methods proposed by Makela, Si and Gelman (2014) for assessing weight adjustments in surveys can be used with IPW as well. Brick (2015) suggested looking at the magnitude of the IPW adjustments in the weighting cells. One can also examine the distribution of the weights within domains of interest.

The inverse propensity weights can also provide information about assumption (A2). A domain that has high weights relative to other domains may have undercoverage in S_A. Dever (2018) proposed investigating assumption (A2) by identifying individuals in S_B who have no close match in S_A.

Bondarenko and Raghunathan (2016) reviewed and proposed graphical and numerical diagnostic tools for assessing and improving imputation models. None of these diagnostics, however, will test the assumption that the regression model fit on S_A applies to units not in S_A. Just as $\bar{y}_A$ may be a biased estimator of μ_y, regression coefficients derived from S_A may also be biased, and the model constructed from S_A to predict y from $\mathbf{x}$ might not apply to other parts of the population.

Take a small probability sample to investigate assumptions

The preceding steps can identify some model deficiencies, but cannot fully test assumptions (A1) and (A2). But one can test the imputation model by obtaining data about y on a probability subsample of S_B. Similarly, one could take a probability sample from population members not in S_A to check inferences from the IPW approach, or observe y on a subsample of units in S_B that are similar to those with high weights in S_A, or that have no close match in S_A.

3. When should one use nonprobability samples?

Wu describes methods for combining information from probability and nonprobability samples after the decision has been made to do so. A first question, however, is whether the operation should be done at all. It may be desired to use a nonprobability sample because no high-quality probability sample measures y, and it is thought that "any information is better than no information". But is that true?

Suppose that, despite the careful model-fitting and model-checking, key statistics are still biased. Could reporting a flawed statistic be worse than reporting no statistic? Bad statistics, once published, can circulate for a long time – even after more rigorous studies show that they are biased. In 1975, advice columnist Ann Landers asked her readers to respond to the question "If you had it to do over again, would you have children?" About 70% of the 10,000 persons who mailed a response said they would not have children in a do-over. This statistic is still cited, even though it is from a convenience sample, has been contradicted by numerous other studies, and is nearly 50 years old (Lohr, 2022). It is also unlikely that predictive modeling or IPW would have corrected the selection bias affecting Landers' statistic, which occurred within all demographic groups.

With these issues in mind, here are some questions that could be asked when deciding whether to use estimates from a nonprobability sample and, if so, which statistical method to use for making inferences.

- How will the statistics be used? Estimates from the nonprobability sample might serve well for developing a marketing strategy or for an exploratory sociological study, but might not be deemed reliable enough for estimating unemployment or the number of persons requiring food assistance. Statistics from a nonprobability sample should be accompanied by evidence that the estimates are fit for use.
- What is the quality of the data in S_A? Administrative records such as tax records have a different quality profile than a survey of volunteers recruited through an internet advertisement.

  If the population for S_A is well-defined (for example, tax filers), it may be better to report statistics for that population than to attempt to generalize to the population of S_B. For tax records, many persons below preset income thresholds have $\pi_i^A = 0$ and assumption (A2) is violated. Instead, a multiple-frame approach might be adopted, where a different data source is used to estimate μ_y for the parts of the population not in S_A (Lohr, 2021).

  Since all of the models rely on auxiliary information $\mathbf{x}$, it is important to have S_A and S_B measure the $\mathbf{x}$ variables the same way. If income is used as an auxiliary variable, the same questions should be used to define income in both surveys, and income should be measured for the same unit (person or household).

  Kennedy (2022) suggested that some respondents to opt-in online surveys may provide incorrect demographic information or bogus answers to questions; if that occurs, model predictions will be flawed. It may even be possible for outsiders desiring a specific outcome to manipulate the data in S_A – for example, an organization might arrange for the survey to be taken by a set of volunteers whose claimed demographic characteristics match those of the population but who give the "desired" answer for y. Some proponents of nonprobability samples argue that low-response-rate probability samples also require weighting adjustments or imputation, but there is one important difference: the probability survey may have nonresponse, but the initial sample is selected randomly and cannot be manipulated by outside organizations.

  If the data in S_A are low-quality, is it worth spending the time to construct models? As Louis (2016) said, "Space-age procedures will not rescue stone-age data".

- How detailed is the auxiliary information? If S_A is large, and the auxiliary information is specific enough to be able to identify specific records, then linking records between S_A and S_B would be a better method for combining the data. Imputation or IPW would be used if the auxiliary information $\mathbf{x}$ is rich enough to give good predictions of y_i or π_i^A, but not rich enough to permit accurate linkage. If there is little auxiliary information, however, then one would expect low variation in the propensity scores or imputed values, and the methods may give poor predictions – with little information to diagnose potential problems.
- What analyses are desired? Wu discusses estimating the population mean, but the analyst may also want to look at relationships between y and other variables, or estimate means or medians

for subgroups. The choice of method depends in part on the variables that are available in S_A and S_B. If S_A contains many response variables whose relationship is of interest, the IPW approach might be preferred.

If it is desired to explore relationships between y and variables measured only in S_B, imputation might be a better choice. Here, though, the analyst should be careful to acknowledge the imputation when presenting results – if, say, linear regression is used for the imputation, the correlation calculated on S_B is not between variable u and variable y, but between u and $\mathbf{x}^T\hat{\boldsymbol{\beta}}$.

- What are the implications for data equity? Jagadish, Stoyanovich and Howe (2021) defined "representation equity" as "increasing the visibility of underrepresented groups that have been historically disadvantaged or suppressed in the data record".

  Nonprobability samples have the potential to improve data equity. They can increase the sample size and visibility of rare population subgroups – a large data set S_A might contain 10,000 members of the subgroup, while even a full-response probability survey with $n_B = 60{,}000$ might contain only ten. Or the nonprobability sample may contain population members who are underrepresented in the probability survey because they are out of scope, undercovered in the sampling frame, or prone to nonresponse. In these situations, S_A provides information about groups that are not as well represented in the probability survey.

  On the other hand, historically disadvantaged groups may be underrepresented in all data sources, including S_A. For example, a large nonprobability sample of electronic health records will be able to generate estimates for more population subgroups than a small probability sample about health. But persons without health insurance or access to medical care are underrepresented. In this situation, relying on S_A to produce population estimates may reinforce inequities. If the estimates are used to distribute resources, then, as the program is implemented, more data will be collected in the areas getting those resources and will validate their needs, but no such follow-up will be done in areas that are inaccurately determined to receive no resources. The feedback loop will propagate the inequitable representation in data sources.

  The MI and IPW methods have different data equity implications. Imputation assigns a predicted value of y to each observation in S_B, and the imputed y value may differ from the y value the respondent would have supplied if asked – particularly if the respondent is in a subgroup that is unrepresented or misrepresented in S_A. Will the model give accurate predictions for historically underrepresented subgroups? Did the respondents to S_B give informed consent for y to be imputed?

  IPW assumes that the propensity scores can be estimated from auxiliary information. Is that information rich enough to give accurate weights? Are some subgroups unrepresented in S_A? It

may be useful to compare the results from the two methods, and from other data sources if available, for historically underrepresented population subgroups.

Wu's critical review raises many important issues for persons interested in using nonprobability samples to make inferences about the population. I especially appreciate his assessment of the strong assumptions needed for the model-based methods, and applaud the emphasis on addressing these problems during the survey design stage.

References

Bondarenko, I., and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, 35(17), 3007-3020.

Brick, J.M. (2015). Compositional model inference. In *Proceedings of the Survey Research Methods Section*, Alexandria, VA: American Statistical Association, 299-307.

Chipperfield, J., Chessman, J. and Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, 54, 223-238.

Dever, J.A. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In *Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference*. https://nces.ed.gov/FCSM/pdf/A4_Dever_2018FCSM.pdf.

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81(S1), 213-239.

Jagadish, H.V., Stoyanovich, J. and Howe, B. (2021). COVID–19 brings data equity challenges to the fore. *Digital Government: Research and Practice*, 2(2), 1-7.

Kennedy, C. (2022). Exploring the assumption that online opt-in respondents are answering in good faith. Paper presented at the 2022 Morris Hansen Lecture, March 1, 2022.

Kim, J.-K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, 47, 2, 229-263. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00008-eng.pdf.

Lohr, S.L. (2022). *Sampling: Design and Analysis, Third Edition*. Boca Raton, FL: CRC Press.

Louis, T.A. (2016). Discussion of combining information from survey and non-survey data sources: Challenges and opportunities. 130th CNSTAT Meeting Public Seminar; Washington, DC. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_172505.pdf.

Makela, S., Si, Y. and Gelman, A. (2014). Statistical graphics for survey weights. *Revista Colombiana de Estadística*, 37(2), 285-295.

Mercer, A., Lau, A. and Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Washington, DC: Pew Research.