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

Survey Methodology

Author's response to comments on "Statistical inference with non-probability survey samples"

by Changbao Wu

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. 367-373
Statistics Canada, Catalogue No. 12-001-X

Author's response to comments on "Statistical inference with non-probability survey samples"

Changbao Wu[1]

Abstract

This response contains additional remarks on a few selected issues raised by the discussants.

Key Words: Data defect correlation; Double robustness; Inverse probability weighting; Model assumptions; Model-based prediction; Validation sample.

Let me start by thanking the Editor of *Survey Methodology*, Jean-François Beaumont, for organizing the discussions and putting together a glamour array of discussants. Each discussant looked at the topic of non-probability survey samples, and more generally topics on data integration and combining data from multiple sources, with some unique perspectives. I have enjoyed reading the discussions and I believe they are significant contributions to dealing with non-probability and other types of samples with selection bias. In what follows, I will make some additional remarks on a few selected issues raised by the discussants.

Michael A. Bailey

Dr. Bailey focused on the limitations of the estimation methods I presented under the assumptions A1-A4, and called for further development when these assumptions, and the so-called "MAR assumption" A1 in particular, are violated. Bailey used non-probabilistic polling as an example to argue that "non-response (can indeed) depends on the study variable" and the danger of A1 being violated is real.

While the criticism on the limitations of the methods reviewed in my paper is fair and square, the statements "(Wu) is fishing in one fairly specific corner of the pond" and "shying away from MNAR models" seem to show significant underappreciation on the importance of methodological development under the standard assumptions A1-A4 which were used by several authors on non-probability survey samples. First of all, the assumption A1 is on the participation (or inclusion/selection) mechanism for non-probability samples, which is not the same as "non-response". There are many scenarios where these assumptions can indeed be justified, especially for surveys using web- or phone-panels where the initial participation in those panels depends largely on certain demographic variables. Second, participation behaviour in non-probability surveys can be confounded by certain study variables during data collection in the same way we face in probability surveys on non-response, which is exactly how the current literature on non-probability surveys has been evolving in dealing with those issues. Third, any methodological advances in addressing the so-called "MNAR models" for non-probability surveys would require the foundation and thorough understanding established under the assumptions A1-A4.

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1.E-mail: cbwu@uwaterloo.ca.

Bailey also stated that "while MAR violations are a problem in probability sampling (arising due to non-response among randomly contacted individuals), MAR violations are more serious in a non-probability world". I heartily concur. As a matter of fact, violations of the positivity assumption A2 are as serious as violations of the "MAR assumption" A1, and the two are intercorrelated. Violations of A2 imply that $\pi_i^A = P(i \in S_A \mid \mathbf{x}_i, y_i) = 0$ for some units in the target population, leading to the undercoverage problem that is as notorious as non-response. When A2 is violated but A1 holds, it is often believed that model-based prediction estimators can mitigate the biases due to undercoverage. Under the assumption A1 the sample inclusion indicator variable R and the study variable y are conditionally independent given $\mathbf{x}$, which implies that

$$E(y_i \mid \mathbf{x}_i, R_i = 1) = E(y_i \mid \mathbf{x}_i). \tag{1}$$

It follows that a valid prediction model $y \mid \mathbf{x}$ can be built using the observed data $\{(y_i, \mathbf{x}_i), i \in S_A\}$ (i.e., units with $R_i = 1$). Unfortunately, the equation (1) implicitly requires $P(R_i = 1 \mid \mathbf{x}_i) > 0$, and prediction-based estimators are not immune to potential undercoverage biases. Bailey's call for "a framework that encompasses the possibility of MAR violations" is in line with some of the current research effort on dealing with undercoverage and "non-ignorable" participation mechanisms for non-probability survey samples. See, for instance, Chen, Li and Wu (2023), Cho, Kim and Qiu (2022) and Yuan, Li and Wu (2022), among others. In a nutshell, valid statistical inferences under those scenarios require either external data such as a validation sample or additional assumptions such as the existence of instrumental variables.

I am on the exact same page of discontent as Bailey with the "missing at random" label, since the term might be confused with "randomly missing" (Wu and Thompson, 2020, page 195). The term "ignorable" is also an unfortunate choice of terminology for missing data and causal inference literature, since it certainly cannot be ignored by the data analyst (Rivers, 2007). I use the standard term "propensity scores" for non-probability samples, while several other authors are in favour of "participation probabilities", including Beaumont (2020) and Rao (2021).

Michael R. Elliott

Dr. Elliott discussed several issues with augmented materials and an expanded list of references. They are important additions to the current topic, especially the reviews on "additional approaches to combining data from probability and non-probability surveys" and sensitivity analysis on "unverifiable assumptions".

Elliott's discussions on distinctions between descriptive parameters and analytic parameters and weighting versus modelling raised the critical issue of efficiency of the IPW estimators in practice. It has been known for probability survey samples that the inverse probability weighted Horvitz-Thompson estimator of the population total T_y is extremely inefficient (in terms of large variance) when the sample selection probabilities π_i are unequal but have very weak correlation to the study variable y, although the estimator remains unbiased under such scenarios. Basu's elephant example (Basu, 1971) described a

"convincing case" where the inverse probability weighted and unbiased Horvitz-Thompson estimator failed miserably, leading to the dismissal of the circus statistician. Discussions on weighting versus modelling, i.e., the IPW estimators versus model-based prediction estimators for descriptive population parameters, are highly relevant for both theoretical developments and practical applications. Our job as a statistician in dealing with non-probability survey samples could be very much in limbo unless we develop solid guidelines and diagnostic tools for choosing suitable approaches with the given dataset and inferential problems.

Elliott echoed my call for a few large scale probability surveys with rich information on auxiliary variables with the statement "it is increasingly critical for an organized and ideally government funded stable of high-quality probability surveys to be put into place for routine data collection". His comments on new areas of research on issues with privacy and confidentiality due to the need for microdata under the context of analyzing non-probability survey samples are a visionary call and deserve an increased amount of attention from the research community.

Zhonglei Wang and Jae Kwang Kim

Dr. Wang and Dr. Kim presented two new approaches to propensity score based estimation, one uses the so-called information projection through a density ratio model and the other employs uniformly calibration functions over a reproducing kernel Hilbert space. These are new adventures in the field, and Kim and his collaborators have the experience and the analytic power to move the research forward.

The starting point for both approaches is the following equation which connects the propensity scores to the density ratios,

$$\frac{1}{P(R_i=1\,|\,\mathbf{x}_i, y_i)} = 1 + \frac{P(R_i=0)}{P(R_i=1)}\frac{f_0(\mathbf{x}_i, y_i)}{f_1(\mathbf{x}_i, y_i)}.$$

The propensity scores $\pi_i^A = P(R_i=1\,|\,\mathbf{x}_i, y_i)$ only require the model on $R_i=1$ given $\mathbf{x}_i$ and y_i. Justification of the equation given above, however, requires a joint randomization framework involving both the model q for the propensity scores and the superpopulation model ξ on $(\mathbf{x}, y)$. From a consistency view point regarding the final estimator of the finite population mean of y, the joint framework imposes very little restrictions if the density ratios are modelled nonparametrically. The consequential impact of the approach is on variance and variance estimation. Variance of an estimator under a joint randomization framework involves more than one component, and variance estimation has further complications if nonparametric procedures are involved. Efficiency comparisons between the proposed methods and some of the existing methods need to be carried out under suitable settings. I am eager to see further developments on the proposed methods.

Sharon L. Lohr

Dr. Lohr's extended discussions on diagnostic tools for assessing model assumptions are highly valuable to the topic. Her explorations of existing ideas and methods and the adaptations to the current

setting highlight the seemingly different but deeply connected issues faced by both nonprobability and probability survey samples. One such issue is the undercoverage problem (i.e., violations of assumption A2) and the interweave of assumptions A1 and A2. Lohr was rightfully concerned with prediction based estimators where the prediction model of y given $\mathbf{x}$ is built based on the nonprobability sample S_A and the mass imputation estimator is computed using observed $\mathbf{x}$ in the reference probability sample S_B, a scenario where each of the two assumptions A1 and A2 does not stand alone. The undercoverage problem is an example where "space-age procedures will not rescue stone-age data". Lohr advocated to "take a small probability sample to investigate assumptions", which is of necessity in theory since rigorously defendable methods under certain scenarios require validation samples. Developments of compromising strategies with existing data sources, however, are more appealing but also more challenging in practice.

Lohr's observation "nonprobability samples have the potential to improve data equity" is an important one, since inclusion of units from groups which may be invisible in probability samples can be boosted relatively easily for nonprobability samples. Lohr also observed that "historically disadvantaged groups may be underrepresented in all data sources, including (nonprobability samples)". Addressing the issue of data equity with nonprobability survey samples presents both opportunities and challenges.

Lohr's question "when should one use nonprobability samples" is a tough one. The same question can be asked for any other statistical methods. We do not seem to always question the validity of the methods and the usefulness of the results in many other scenarios due to our unchecked confidence that the required assumptions seem to be reasonable. For nonprobability samples, we have a more vulnerable situation regarding assumptions, and assessments and diagnostics of these assumptions are more difficult than cases with controlled experiments and/or more structured data. From this view point, Lohr's extended discussion on assessing assumptions should be read with deep appreciation. In practice, an important confidence booster on the assumptions is the thorough investigation at the "design stage", if such a stage can be conceived prior to data collection, on variables which might be related to participation behaviour, and to include these variables as part of the sample with further exploration of existing data sources containing these variables.

Xiao-Li Meng

Dr. Meng's discussion, with the formal title "Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples", should be a standalone discussion paper itself. Meng weaved through a number of issues in estimating a finite population mean with a nonprobability sample, and explored strategies and directions for constructing an approximately unbiased estimator using the central concept of the so-called *data defect correlation (ddc)*. The discussions are fascinating and thought-provoking, and will surely generate more discussions and research endeavours on implications of the *ddc*. I would like to use this opportunity to comment briefly on the *ddc* in relation to three basic concepts in probability sampling: *sampling strategy*, *undercoverage*, and *model-assisted estimation*. It is not a nostalgia for the good old days when probability sampling was the golden standard but rather an

appreciation of how research in survey sampling has been evolving and the potential usefulness of the *ddc* in dealing with nonprobability survey samples.

The term *sampling strategy* refers to the pair of sampling design and estimation method (Thompson, 1997, Section 2.4; Rao, 2005, Section 3.1). The two components go hand in hand and are the backbone of conventional probability survey sampling theory. For the estimation of the population total T_y of the study variable y using a probabiliity sample S with first order inclusion probabilities π_i, the Horvitz-Thompson estimator $\hat{T}_{y\text{HT}} = \sum_{i\in S} d_i y_i$ with the weight $d_i = \pi_i^{-1}$ is the unique unbiased estimator among a class of linear estimators (Wu and Thompson, 2020). The theoretical argument for the result is straightforward due to the known inclusion probabilities π_i under the given sampling design. Using the notation of Meng, the *ddc* involves three variables: the study variable G, the weight variable W, the sample inclusion indicator R, and is defined as the finite population correlation coefficient between $\tilde{R} = RW$ and G. The *ddc* implicitly puts R and W as an inseparable pair for any *inference strategy*, with R corresponding to the unknown "design" and W for the "estimation method". With the unknown "design" characterized by the unknown "divine probabilities" π_I for the nonprobability sample, Meng showed through his equation (3.3) that $W_I \propto \pi_I^{-1}$ is essentially a required condition for unbiased estimation of $\bar{G}$ if nothing is assumed on the outcome regression model. The result provides a justification of the use of inverse probability weighted (IPW) estimator for nonprobability samples as the only sensible choice if a superpopulation model on the study variable is not involved.

The problem of *undercoverage* has been discussed extensively in the existing literature on probability sampling. For nonprobability samples the issue is closely related to the violation of the positivity assumption A2 as discussed in Section 7.2 of my paper and my comments to the discussions of Bailey, Elliott and Lohr, with additional details given in Chen et al. (2023). Let $U = U_0 \cup U_1$, where U_1 is the uncovered subpopulation with $\pi_i^A = P(R_i = 1 \,|\, \mathbf{x}_i, y_i) = 0$. Let $N = N_0 + N_1$, where N_0 and N_1 are respectively the sizes of the two subpopulations U_0 and U_1. Let Cov_I and $\text{Cov}_I^{(0)}$ denote respectively the covariance with respect to the discrete uniform distribution over U and U_0. It can be shown that

$$\text{Cov}_I(\tilde{R}_I, G_I) = \omega_0 \left\{ \text{Cov}_I^{(0)}(\tilde{R}_I, G_I) - \omega_1 (\bar{G}_1 - \bar{G}_0)\, \hat{N}_0 / N_0 \right\}, \tag{2}$$

where $\omega_k = N_k / N$ for $k = 0, 1$, $\hat{N}_0 = \sum_{i\in S} W_i$, S is the set of units for the nonprobability sample, and $\bar{G}_0$ and $\bar{G}_1$ are respectively the population means of U_0 and U_1 for the study variable G. Equation (2) has two immediate implications. First, if the estimation method is valid in the sense that the value of $\text{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ is small, then the bias of the estimator $\bar{G}_W$ due to undercoverage depends on ω_1 (i.e., the size of the uncovered subpopulation U_1) and $\bar{G}_1 - \bar{G}_0$ (i.e., the difference between U_0 and U_1), a statement which has previously been established under probability sampling. Second, the equation reveals a scenario for potential *counterbalancing*: A biased estimator $\bar{G}_W$ for the "sampled population mean" $\bar{G}_0$ can be less biased for the target population mean $\bar{G}$ if $\text{Cov}_I^{(0)}(\tilde{R}_I, G_I)$ and $\bar{G}_1 - \bar{G}_0$ have the same plus or minus sign.

Meng's discussions on quasi-randomization and/or super-population using the *ddc* provided a much deeper understanding on doubly robust estimation. Historically, *model-assisted estimation* started to emerge in survey sampling in the early 1970s, and the approach has the same spirit of double robustness. The generalized difference estimator of the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ as discussed in Cassel, Särndal and Wretman (1976) is given by

$$\hat{\mu}_{y\text{GD}} = \frac{1}{N}\left\{\sum_{i\in S}\frac{y_i - c_i}{\pi_i} + \sum_{i=1}^{N} c_i\right\}, \tag{3}$$

where S is a probability sample, the π_i's are the first order inclusion probabilities, and $\{c_1, c_2, \ldots, c_N\}$ is an arbitrary sequence of known numbers. The estimator $\hat{\mu}_{y\text{GD}}$ is exactly unbiased for μ_y under the probability sampling design p for any sequence c_i, and is also model-unbiased if we choose $c_i = m_i = E_\xi(y_i \mid \mathbf{x}_i)$. Cassel et al. (1976) showed a main theoretical result that the choice $c_i = m_i$ is optimal leading to minimum model-based expectation of the design-based variance $E_\xi\{V_p(\hat{\mu}_{y\text{GD}})\}$ when the model has certain structure in variance. The first part of the results on unbiasedness is under (p or ξ); the second part on optimality is under (p and ξ). Note that the estimator $\hat{\mu}_{y\text{GD}}$ with the choice $c_i = \hat{m}_i$ has exactly the same structure of the doubly robust estimator discussed extensively in the missing data and causal inference literature since the 1990s, with the "divine probabilities" π_i being unknown and estimated in the latter cases.

The use of *ddc* in practice requires additional information from the population. Meng's proposal of creating a representative miniature out of a biased sample echoes the call for a validation sample with a small size, since such a sample "can (also) eliminate many practitioners's anxiety and potential mistakes for not knowing how to properly use the weights".

"There is no such thing as probability sample in real life" is probably a defendable statement for human populations. Probability samples, however, do exist in other fields such as business and establishment surveys, agricultural surveys, and natural resource inventory surveys; see Wu and Thompson (2020) for further detail. For humans, any rigorous rules and precise procedures are *almost surely* as aspiration, not prescription.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling. Part One. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto, 203-242.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chen, Y., Li, P. and Wu, C. (2023). Dealing with undercoverage for non-probability survey samples. *Survey Methodology*, under review.

Cho, S., Kim, J.K. and Qiu, Y. (2022). *Multiple Bias Calibration for Valid Statistical Inference with Selection Bias*. Working paper.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivers, D. (2007). Sampling for web surveys. In *Proceedings of the Survey Research Methods Section*, Joint Statistical Meetings, American Statistical Association, Alexandria, VA, 1-26.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, London.

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Springer, Cham.

Yuan, M., Li, P. and Wu, C. (2022). *Inference with Non-Ignorable Sample Inclusion for Non-Probability Survey Samples*. Working paper.