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

Survey Methodology

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. 283-311
Statistics Canada, Catalogue No. 12-001-X

Statistical inference with non-probability survey samples

Changbao Wu[1]

Abstract

We provide a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. We attempt to present rigorous inferential frameworks and valid statistical procedures under commonly used assumptions, and address issues on the justification and verification of assumptions in practical applications. Some current methodological developments are showcased, and problems which require further investigation are mentioned. While the focus of the paper is on non-probability samples, the essential role of probability survey samples with rich and relevant information on auxiliary variables is highlighted.

Key Words: Auxiliary information; Bootstrap variance estimator; Calibration method; Doubly robust estimator; Estimating equations; Inverse probability weighting; Model-based prediction; Poststratification; Pseudo likelihood; Propensity score; Quota survey; Sensitivity analysis; Variance estimation.

1. Introduction

The field of survey sampling distinguishes itself from other areas of statistics with a number of unique features. The target population consists of finite number of well defined units, and the population parameters can be determined without error, at least conceptually, by conducting a census. Operational constraints and administrative convenience for data collection often make it necessary to consider stratification, clustering and unequal probability selection. Since the seminal paper of Neyman (1934), probability sampling methods have become one of the primary data collection tools for official statistics and researchers in health sciences, social and economic studies, business and marketing, agricultural and natural resource inventories, and other areas. Probability survey samples have also been used for analytic studies involving models and model parameters; see, for instance, Binder (1983), Godambe and Thompson (1986), Thompson (1997), Rao and Molina (2015), among others. Probability survey samples and design-based inference have been a successful story as part of statistical sciences in the past 80 years.

In recent years, however, "*there has been a wind of change and other data sources are being increasingly explored*" (Beaumont, 2020). The success of probability survey samples led to more ambitious study designs, long and complicated questionnaires and increased burden on respondents. The response rates have been declining and the cost of data collection has been soaring over the years. With the advances of new technology and the explosion of information over the Internet, there is also a strong desire to access real-time statistics. Statistics Canada has launched the so-called modernization initiatives, "*moving beyond a survey-first approach with new methods and integrating data from a variety of existing sources*".

Non-probability survey samples are one of those data sources which have gained increased popularity in recent years. Non-probability samples are not something new to the field of survey sampling. They have been used since the early days of conducting surveys. Quota surveys, for instance, lead to

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

non-probability samples, and the method is widely used and can be successful under certain conditions; see Section 5 for further discussions. Non-probability survey samples had not gained true momentum in the past in survey practice due to the lack of a mature theoretical framework for analyzing the data. Nevertheless, they are an available data source that is cheaper and quicker to obtain and have become prevalent for online research. Commercial survey firms create and maintain a long list of individuals, called the *opt-in panels*, who agreed to be contacted to participate in surveys either as volunteers or with incentives. The precise mechanisms for individuals being included in the panel are typically unknown, resulting in panel-based non-probability survey samples.

The main issue with non-probability survey samples is that they are biased samples and do not represent the target population. One might argue that, other than iid samples, most samples are biased, and even probability survey samples are biased. The reason that we do not worry about the biased nature of probability survey samples is the known inclusion probabilities from the survey design, which lead to valid estimation methods through suitable weighting procedures. The real main issue with non-probability survey samples thus is the unknown sample inclusion or participation mechanisms. It will become clear from discussions in Section 4 that the biased nature of non-probability samples cannot be corrected by using the sample itself. It requires additional auxiliary information on the target population.

This paper provides a critical review and some extended discussions on theoretical and practical issues with analysis of non-probability survey samples. Section 2 describes the general setting, commonly used assumptions, and inferential frameworks for statistical procedures discussed in the paper. Section 3 presents model-based prediction approach to non-probability survey samples. Section 4 discusses estimation of propensity scores and constructions of propensity score based estimators. Section 5 shows the connections between inverse probability weighted estimators and quota surveys with extensions to poststratification. Section 6 focuses on techniques as well as issues with variance estimation. In Section 7, we address the important question on how to check and verify the required assumptions in practice. Some concluding remarks are given in Section 8.

2. Assumptions and inferential frameworks

Suppose that the target population $U = \{1, 2, \ldots, N\}$ consists of N labelled units. Associated with unit i are values $\mathbf{x}_i$ and y_i for the auxiliary variables $\mathbf{x}$ and the study variable y. The discussions focus on a single y but the dataset most likely contains multiple study variables. Let $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ be the population mean which is the parameter of interest. Let $\{(y_i, \mathbf{x}_i), i \in S_A\}$ be the dataset for the non-probability survey sample S_A with n_A participating units. For most practical scenarios, the simple sample mean $\bar{y}_A = n_A^{-1}\sum_{i \in S_A} y_i$ is a biased estimator of μ_y and hence is invalid.

2.1 Assumptions

Let $R_i = I(i \in S_A)$ be the indicator variable for unit i being included in the non-probability sample S_A. Note that the variable R_i is defined for all i in the target population. Let

$$\pi_i^A = P\left(i \in S_A \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right), \quad i = 1, 2, \ldots, N.$$

We call the π_i^A the propensity scores, a term borrowed from the missing data literature (Rosenbaum and Rubin, 1983). Some authors use the term participation probabilities; see, for instance, Beaumont (2020) and Rao (2021), among others. The propensity scores π_i^A characterize the sample inclusion and participation mechanisms. They are unknown and require suitable model assumptions for the development of valid estimation methods. The following three basic assumptions were used by Chen, Li, and Wu (2020), which were adapted from the missing data literature.

A1 The sample inclusion and participation indicator R_i and the study variable y_i are independent given the set of covariates $\mathbf{x}_i$, i.e., $\left(R_i \perp y_i\right) \middle| \mathbf{x}_i$.

A2 All the units in the target population have non-zero propensity scores, i.e., $\pi_i^A > 0$, $i = 1, 2, \ldots, N$.

A3 The indicator variables $R_1, R_2, \ldots, R_N$ are independent given the set of auxiliary variables $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$.

Assumption A1 is similar to the missing at random (MAR) assumption for missing data analysis. Under A1, we have $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$. Assumption A2 can be problematic in practice; see Section 7 for further discussions. Assumption A3 typical holds when participants are approached one at a time but can be questionable when clustered selections are used. It is shown in Section 4 that estimation of $\pi_i^A = \pi\left(\mathbf{x}_i\right)$ under assumption A1 requires auxiliary information from the target population. The ideal scenario is that the complete auxiliary information $\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right)$ is available. The more practical scenario is that auxiliary information can be obtained from an existing probability survey.

A4 There exists a probability survey sample S_B of size n_B with information on the auxiliary variables $\mathbf{x}$ (but not on y) available in the dataset $\left\{(\mathbf{x}_i, d_i^B), i \in S_B\right\}$, where d_i^B are the design weights for the probability sample S_B.

The S_B is called the reference probability survey sample. The most crucial part of assumption A4 is that the set of auxiliary variables $\mathbf{x}$ is observed in both the non-probability sample S_A and the probability sample S_B. A reference probability survey sample is often available in practice but the common set of auxiliary variables may not contain all the components to satisfy assumption A1.

2.2 Inferential frameworks

There are three possible sources of variation under the general setting of two samples S_A and S_B: (i) The model q for the propensity scores on the sample inclusion and participation in the non-probability survey sample S_A; (ii) The model ξ for the outcome regression $\left(y \middle| \mathbf{x}\right)$ or imputation; and (iii) The probability sampling design p for the reference probability survey sample S_B. For the three approaches

to inference to be discussed in Sections 3 and 4, the reference probability sample S_B is always involved. Each of the three approaches requires a joint randomization framework involving p and one of (q, ξ).

(a) Model-based prediction approach: The ξp framework under the joint randomization of the outcome regression model ξ and the probability sampling design p.

(b) Inverse probability weighting using estimated propensity scores: The qp framework under the joint randomization of the propensity score model q and the probability sampling design p.

(c) Doubly robust inference: The qp framework or the ξp framework, with no specification of which one.

The inferential framework is the foundation for theoretical development. Consistency of point estimators needs to be established under the suitable joint randomization. Theoretical variances typically involve two components, one from each source of variation, and correct derivations of the two components are the key to the construction of consistent variance estimators under the designated inferential framework.

3. Model-based prediction approach

Model-based prediction methods for finite population parameters require two critical ingredients: the amount of auxiliary information that is available at the estimation stage and the reliability of the assumed model for inference. In the absence of any auxiliary information, the common mean model $E_\xi(y_i) = \mu_0$, $V_\xi(y_i) = \sigma^2$, $i = 1, \ldots, N$ may be viewed as reasonable but the model-based prediction estimator $\hat{\mu}_y = \bar{y}_A = n_A^{-1} \sum_{i \in S_A} y_i$, although unbiased under the model since $E_\xi(\bar{y}_A - \mu_y) = 0$, is generally not an acceptable estimator of μ_y. The variance σ^2 for the common mean model is typically large and it renders the estimator $\hat{\mu}_y = \bar{y}_A$ with a prediction variance that is too large to be practically useful.

3.1 Semiparametric outcome regression models

Without loss of generality, we assume that $\mathbf{x}$ contains 1 as its first component corresponding to the intercept of a regression model. Under the setting described in Section 2, we consider the following semiparametric model for the finite population, denoted as ξ:

$$E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta}), \quad \text{and} \quad V_\xi(y_i \mid \mathbf{x}_i) = v(\mathbf{x}_i)\sigma^2, \quad i = 1, 2, \ldots, N, \tag{3.1}$$

where the mean function $m(\cdot, \cdot)$ and the variance function $v(\cdot)$ have known forms, and the y_i's are also assumed to be conditionally independent given the $\mathbf{x}_i$'s. Let $\boldsymbol{\beta}_0$ and σ_0^2 be the true values of the model parameters $\boldsymbol{\beta}$ and σ^2 under the assumed model. The first major implication of assumption A1 is that $E_\xi(y_i \mid \mathbf{x}_i, R_i = 1) = E_\xi(y_i \mid \mathbf{x}_i)$ and $V_\xi(y_i \mid \mathbf{x}_i, R_i = 1) = V_\xi(y_i \mid \mathbf{x}_i)$. The model (3.1) which is assumed for the finite population also holds for the units in the non-probability survey sample S_A. The quasi maximum

likelihood estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ is obtained using the dataset $\{(y_i, \mathbf{x}_i), i \in S_A\}$ from the non-probability survey sample as the solution to the quasi score equations (McCullagh and Nelder, 1989) given by

$$\mathrm{S}(\boldsymbol{\beta}) = \sum_{i \in S_A} \frac{\partial m(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \{v(\mathbf{x}_i)\}^{-1} \{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})\} = \mathbf{0}. \tag{3.2}$$

The semiparametric model (3.1) can be extended to replace $v(\mathbf{x}_i)$ by a general variance function $v(\mu_i)$ where $\mu_i = m(\mathbf{x}_i, \boldsymbol{\beta})$. The quasi maximum likelihood estimation theory covers linear or nonlinear regression models with the weighted least square estimators, the logistic regression model and other generalized linear models. Let $m_i = m(\mathbf{x}_i, \boldsymbol{\beta}_0)$ and $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, $i = 1, 2, \ldots, N$.

3.2 Two general forms of prediction estimators

There are two commonly used model-based prediction estimators for μ_y in the presence of complete auxiliary information $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$; see Chapter 5 of Wu and Thompson (2020). Note that $E_\xi(\mu_y) = N^{-1} \sum_{i=1}^N m_i$. The two prediction estimators are constructed as

$$\hat{\mu}_{y1} = \frac{1}{N} \sum_{i=1}^N \hat{m}_i \quad \text{and} \quad \hat{\mu}_{y2} = \frac{1}{N} \left\{ \sum_{i \in S_A} y_i - \sum_{i \in S_A} \hat{m}_i + \sum_{i=1}^N \hat{m}_i \right\}. \tag{3.3}$$

The estimator $\hat{\mu}_{y2}$ is built based on $\mu_y = N^{-1} \left\{ \sum_{i \in S_A} y_i + \sum_{i \notin S_A} y_i \right\}$ and uses $\sum_{i \notin S_A} \hat{m}_i = \sum_{i=1}^N \hat{m}_i - \sum_{i \in S_A} \hat{m}_i$ to predict the unobserved term $\sum_{i \notin S_A} y_i$. Under a linear regression model where $m(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$, the two estimators given in (3.3) reduce to

$$\hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}' \hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\mu}_{y2} = \frac{n_A}{N} \left(\bar{y}_A - \bar{\mathbf{x}}_A' \hat{\boldsymbol{\beta}} \right) + \boldsymbol{\mu}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}, \tag{3.4}$$

where $\boldsymbol{\mu}_{\mathbf{x}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$ is the vector of the population means of the $\mathbf{x}$ variables and $\bar{\mathbf{x}}_A = n_A^{-1} \sum_{i \in S_A} \mathbf{x}_i$ is the vector of the simple sample means of $\mathbf{x}$ from the non-probability sample S_A. If the linear regression model contains an intercept and $\hat{\boldsymbol{\beta}}$ is the ordinary least square estimator, we have $\hat{\mu}_{y2} = \hat{\mu}_{y1} = \boldsymbol{\mu}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}$ since $\bar{y}_A - \bar{\mathbf{x}}_A' \hat{\boldsymbol{\beta}} = 0$ due to the zero sum of fitted residuals. The prediction estimators in (3.4) under a linear model only require the population means $\boldsymbol{\mu}_{\mathbf{x}}$ in addition to the non-probability sample S_A. Under the setting described in Section 2 with auxiliary information on $\mathbf{x}$ provided through a reference probability sample S_B, we simply replace $\sum_{i=1}^N \hat{m}_i$ by $\sum_{i \in S_B} d_i^B \hat{m}_i$ for the estimators in (3.3) and substitute $\boldsymbol{\mu}_{\mathbf{x}}$ by $\hat{\boldsymbol{\mu}}_{\mathbf{x}} = \hat{N}_B^{-1} \sum_{i \in S_B} d_i^B \mathbf{x}_i$ for the estimators in (3.4), where $\hat{N}_B = \sum_{i \in S_B} d_i^B$. The population size N appearing in (3.3) or (3.4) should also be replaced by $\hat{N}_B$ even if it is known.

3.3 Mass imputation

Model-based prediction estimators of μ_y using a non-probability survey sample on $(y, \mathbf{x})$ and a reference probability survey sample on $\mathbf{x}$ have traditionally been presented as the *mass imputation estimator*. The study variable y is not observed for any units in the reference survey sample S_B and hence

can be viewed as missing for all $i \in S_B$. Let y_i^* be an imputed value for y_i, $i \in S_B$. The mass imputation estimator of μ_y is then constructed as

$$\hat{\mu}_{y\text{MI}} = \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B y_i^*, \tag{3.5}$$

where $\hat{N}_B$ is defined as before and the subscript "MI" indicates "Mass Imputation" (not "Multiple Imputation"). Under the deterministic regression imputation where $y_i^* = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$, the estimator $\hat{\mu}_{y\text{MI}}$ reduces to the model-based prediction estimator $\hat{\mu}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}$ as discussed in Section 3.2.

The mass imputation approach to analyzing non-probability survey samples has the same spirit as model-based prediction methods but it opens the door for using more flexible models and imputation techniques that have been developed in the existing literature on missing data problems. The approach was first examined by Rivers (2007) through the so-called *sample matching* method. For each $i \in S_B$, the "missing" y_i is imputed as $y_i^* = y_j$ for some $j \in S_A$, where j is a matching donor from S_A selected through the nearest neighbor method as measured by the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The underlying model ξ for the nearest neighbor imputation method is nonparametric, i.e., $E_\xi \left(y_i \mid \mathbf{x}_i \right) = m\left(\mathbf{x}_i \right)$ for some unknown function $m(\cdot)$. The matching value y_j can be viewed as the predicted value of the missing y_i under the model. Theoretical properties of estimators based on nearest neighbor imputation were discussed by Chen and Shao (2000, 2001) for missing survey data problems.

The semiparametric model (3.1) can be used for deterministic regression mass imputation. Under assumption A1, a consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is first obtained from the non-probability sample dataset $\left\{ (y_i, \mathbf{x}_i), i \in S_A \right\}$, and the estimator $\hat{\boldsymbol{\beta}}$ is then used to compute the imputed values $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}} \right)$ for $i \in S_B$. In other words, the assumption A1 implies the so-called *model transportability* by Kim, Park, Chen and Wu (2021): the model which is built for the non-probability sample can be used for prediction with the reference probability sample. The resulting mass imputation estimator $\hat{\mu}_{y\text{MI}}$ is identical to one of the model-based prediction estimators presented in Section 3.2. Asymptotic properties and variance estimation for the estimator $\hat{\mu}_{y\text{MI}}$ using the semiparametric model (3.1) were discussed by Kim et al. (2021).

Under the mass imputation approach, the only role played by the observed y_i for $i \in S_A$ is to estimate the model parameters $\boldsymbol{\beta}$. The estimator $\hat{\mu}_{y\text{MI}}$ is constructed using the fitted model and auxiliary information from the reference probability sample S_B. It seems that we did not fully use the information on the observed y_i given that μ_y is the main parameter of interest. This led to the research question described in Chapter 17 of Wu and Thompson (2020) on "*reverse sample matching*". The proposed estimator is constructed as $\hat{\mu}_{y\text{A}} = (\hat{N}^*)^{-1} \sum_{i \in S_A} d_i^* y_i$ using all the observed y_i in the non-probability sample, where $\hat{N}^* = \sum_{i \in S_A} d_i^*$. The d_i^* is a matched survey weight from S_B such that $d_i^* = d_j^B$ with $j \in S_B$ being the nearest neighbor of $i \in S_A$ as measured by $\left\| \mathbf{x}_i - \mathbf{x}_j \right\|$. Theoretical properties of the reverse matched estimator $\hat{\mu}_{y\text{A}}$ using the nearest neighbor $j \in S_B$ to match d_i^* with d_j^B have not been formally investigated in the existing literature.

Wang, Graubard, Katki and Li (2020) proposed a kernel weighting approach to reverse sample matching using $d_i^* \propto \sum_{j \in S_B} K_{ij} d_j$, where K_{ij} is a kernel distance between $\hat{p}_i$ and $\hat{p}_j$; see the adjusted logistic propensity (ALP) weighting method discussed at the end of Section 4.1.1 on the calculation of $\hat{p}_i$. They showed that the estimator $\hat{\mu}_{yA}$ is consistent under certain regularity conditions. In a recent working paper posted on arXiv by Liu and Valliant (2021), the authors discussed issues with the bias and the variance of the reverse matched estimator under different randomization frameworks involving one, two or all three of the sources (p, q, ξ). The authors also proposed a calibration step over the matched weights, which seems to be a promising idea. Further research on this topic is needed.

The mass imputation approach to analyzing non-probability survey samples leads to an interesting research question that is currently under investigation by a doctoral student at University of Waterloo: Is it theoretically feasible and practically useful to create a mass-imputed dataset $\left\{(y_i^*, \mathbf{x}_i, d_i^B), i \in S_B\right\}$ based on the reference probability survey sample that can be used for general statistical inferences? The answer clearly depends on the types of inferential problems to be conducted over the imputed dataset. A minimum requirement is that the conditional distribution of the study variable y given the covariates $\mathbf{x}$ is preserved for the mass-imputed dataset. The nearest neighbor imputation method and the random regression imputation method can be useful for this purpose. Fractional imputation is another possibility, especially for binary or ordinal study variables. Multiple imputation is also potentially useful in this direction to create multiple mass-imputed datasets. The subscript "MI" in this case might need to be changed to "MI²", meaning "Mass Imputation with Multiple Imputation".

4. Propensity scores based approach

The propensity scores $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right)$ for the non-probability survey sample S_A are theoretically defined for all the units in the target population. Estimation of the propensity scores for units in S_A, which plays the most crucial role for propensity scores based methods, requires an assumed model on the propensity scores and auxiliary information at the population level. In this section, we first discuss estimation procedures for the propensity scores under the setting and assumptions described in Section 2, and then provide an overview of estimation methods proposed in the recent literature on the finite population mean μ_y involving the estimated propensity scores.

4.1 Estimation of propensity scores

Under assumption A1, the propensity scores $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i\right) = \pi\left(\mathbf{x}_i\right)$ are a function of the auxiliary variables $\mathbf{x}_i$ but the functional form can be complicated and is completely unknown. Three popular parametric forms $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$ in dealing with a binary response can be considered: (i) the inverse logit function $\pi_i^A = 1 - \left\{1 + \exp\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)\right\}^{-1}$; (ii) the inverse probit function $\pi_i^A = \Phi\left(\mathbf{x}_i'\boldsymbol{\alpha}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$; and (iii) the inverse complementary log-log function

$\pi_i^A = 1 - \exp\{-\exp(\mathbf{x}_i'\boldsymbol{\alpha})\}$. Nonparametric techniques without assuming an explicit functional form for $\pi(\mathbf{x})$ are attractive alternatives for the estimation of propensity scores.

4.1.1 The pseudo maximum likelihood method

Let $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ be a specified parametric form with unknown model parameters $\boldsymbol{\alpha}$. Under the ideal situation where the complete auxiliary information $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is available and with the independence assumption A3, the full log-likelihood function on $\boldsymbol{\alpha}$ can be written as (Chen et al., 2020)

$$\ell(\boldsymbol{\alpha}) = \log\left\{\prod_{i=1}^{N}\left(\pi_i^A\right)^{R_i}\left(1-\pi_i^A\right)^{1-R_i}\right\} = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i=1}^{N}\log\left(1-\pi_i^A\right). \tag{4.1}$$

The maximum likelihood estimator of $\boldsymbol{\alpha}$ is the maximizer of $\ell(\boldsymbol{\alpha})$. Under the current setting where the population auxiliary information is supplied by the reference probability sample S_B, we replace $\ell(\boldsymbol{\alpha})$ by the pseudo log-likelihood function (Chen et al., 2020)

$$\ell^*(\boldsymbol{\alpha}) = \sum_{i\in S_A}\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i\in S_B} d_i^B \log\left(1-\pi_i^A\right). \tag{4.2}$$

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\alpha}}$ is the maximizer of $\ell^*(\boldsymbol{\alpha})$ and can be obtained as the solution to the pseudo score equations given by $\mathbf{U}(\boldsymbol{\alpha}) = \partial\ell^*(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = \mathbf{0}$. If the inverse logit function is assumed for π_i^A, the pseudo score functions are given by

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{i\in S_A}\mathbf{x}_i - \sum_{i\in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha})\,\mathbf{x}_i. \tag{4.3}$$

In general, the pseudo score functions $\mathbf{U}(\boldsymbol{\alpha})$ at the true values of the model parameters $\boldsymbol{\alpha}_0$ are unbiased under the joint qp randomization in the sense that $E_{qp}\{\mathbf{U}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$, which implies that the estimator $\hat{\boldsymbol{\alpha}}$ is qp-consistent for $\boldsymbol{\alpha}_0$ (Tsiatis, 2006).

Valliant and Dever (2011) made an earlier attempt to estimate the propensity scores by pooling the non-probability sample S_A with the reference probability sample S_B. Let $S_{AB} = S_A \cup S_B$ be the pooled sample without removing any potential duplicated units. Let $R_i^* = 1$ if $i \in S_A$ and $R_i^* = 0$ if $i \in S_B$. Valliant and Dever (2011) proposed to fit a survey weighted logistic regression model to the pooled dataset $\{(R_i^*, \mathbf{x}_i, d_i), i \in S_{AB}\}$, where the weights are defined as $d_i = 1$ if $i \in S_A$ and $d_i = d_i^B\left(1 - n_A/\hat{N}_B\right)$ if $i \in S_B$. The key motivation behind the creation of the weights d_i is that the total weight $\sum_{i\in S_{AB}} d_i = \sum_{i\in S_B} d_i^B = \hat{N}_B$ for the pooled sample matches the estimated population size, and the hope is that the survey weighted logistic regression model would lead to valid estimates for the propensity scores. It was shown by Chen et al. (2020) that the pooled sample approach of Valliant and Dever (2011) does not lead to consistent estimators for the parameters of the propensity scores model unless the non-probability sample S_A is a simple random sample from the target population.

The method of Valliant and Dever (2011) reveals a fundamental difficulty with approaches based on the pooled sample S_{AB}. If the units in the non-probability sample S_A are treated as exchangeable in the pooled sample S_{AB}, which was reflected by the equal weights $d_i = 1$ used in the method of Valliant and Dever (2011), the resulting estimates for the propensity scores will be invalid unless S_A is a simple random sample. This observation has implications to the validity of nonparametric methods or regression tree-based methods to be discussed in Section 4.1.3.

In a recent paper, Wang, Valliant and Li (2021) proposed an adjusted logistic propensity (ALP) weighting method. The method involves two steps for computing the estimated propensity scores. The initial estimates, denoted as $\hat{p}_i$ for $i \in S_A$, are obtained by fitting the survey weighted logistic regression model to the pooled sample S_{AB} similar to Valliant and Dever (2011), with the weights defined as $d_i = 1$ if $i \in S_A$ and $d_i = d_i^B$ if $i \in S_B$. The final estimated propensity scores are computed as $\hat{\pi}_i^A = \hat{p}_i / (1 - \hat{p}_i)$. The key theoretical argument is the equation $\pi_i^A = p_i / (1 - p_i)$ where $\pi_i^A = P\left(i \in S_A \mid U\right)$, $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$, and S_A^* is a copy of S_A but is viewed as a different set. However, there are conceptual issues with the arguments since the probabilities $\pi_i^A = P\left(i \in S_A \mid U\right)$ are defined under the assumed propensity scores model with the given finite population U, and the assumed model does not lead to a meaningful interpretation of the probabilities $p_i = P\left(i \in S_A^* \mid S_A^* \cup U\right)$. The latter require a different probability space and are conditional on the given S_A. As a matter of fact, one can easily argue that under the assumed propensity scores model and conditional on the given S_A, we have $p_i = 1$ if $i \in S_A$ and $p_i = 0$ otherwise.

4.1.2 Estimating equations based methods

The pseudo score equations $\mathbf{U}(\boldsymbol{\alpha}) = \mathbf{0}$ derived from the pseudo likelihood function $\ell^*(\boldsymbol{\alpha})$ may be replaced by a system of general estimating equations. Let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ be a user-specified vector of functions with the same dimension of $\boldsymbol{\alpha}$. Let

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - \sum_{i \in S_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha}) \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}). \tag{4.4}$$

It follows that $E_{qp}\{\mathbf{G}(\boldsymbol{\alpha}_0)\} = \mathbf{0}$ for any chosen $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. In principle, an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ can be obtained by solving $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ with the chosen parametric form $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ and the chosen functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$, and the estimator $\hat{\boldsymbol{\alpha}}$ is consistent.

The estimator $\hat{\boldsymbol{\alpha}}$ using arbitrary user-specified functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$ is typically less efficient than the one based on the pseudo score functions, due to the optimality of the maximum likelihood estimator (Godambe, 1960). Some limited empirical results also show that the solution to $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ can be unstable for certain choices of $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. Nevertheless, the estimating equations based methods provide a useful tool for the estimation of the propensity scores under more restricted scenarios. For instance, if we let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha})$, the estimating functions given in (4.4) reduce to

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in S_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \sum_{i \in S_B} d_i^B \mathbf{x}_i. \tag{4.5}$$

The form of $\mathbf{G}(\boldsymbol{\alpha})$ in (4.5) looks like a "distorted" version of the pseudo score functions given in (4.3) under a logistic regression model for the propensity scores. The most practically important difference between the two versions, however, is the fact that the $\mathbf{G}(\boldsymbol{\alpha})$ given in (4.5) only requires the estimated population totals for the auxiliary variables $\mathbf{x}$. There are scenarios where the population totals of the auxiliary variables $\mathbf{x}$ can be accessed or estimated from an existing source but values of $\mathbf{x}$ at the unit level for the entire population or even a probability sample are not available. The use of estimating functions $\mathbf{G}(\boldsymbol{\alpha})$ given (4.5) makes it possible to obtain valid estimates of the propensity scores for units in the non-probability sample. Section 6.3 describes an example where the estimating equations based approach leads to a valid variance estimator for the doubly robust estimator of the population mean.

4.1.3 Nonparametric methods and regression-tree based methods

The propensity scores $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i)$ are the mean function $E_q(R_i \mid \mathbf{x}_i) = \pi(\mathbf{x}_i)$ for the binary response R_i. Nonparametric methods for estimating $\pi(\mathbf{x})$ can be an attractive alternative. The major challenge is to develop estimation procedures which provide valid estimates of the propensity scores. As noted in Section 4.1.1, estimation methods based on the pooled sample $S_{AB} = S_A \cup S_B$ may lead to invalid estimates. Strategies similar to the one used by Chen et al. (2020) can be theoretically justified under the joint qp framework, where the estimation procedures are first derived using data from the entire finite population and unknown population quantities are then replaced by estimates obtained from the reference probability sample.

We consider the kernel regression estimator of $\pi_i^A = \pi(\mathbf{x}_i)$. Suppose that the dataset $\{(R_i, \mathbf{x}_i), i = 1, 2, \ldots, N\}$ is available for the finite population. Let $K_h(t) = K(t/h)$ be a chosen kernel with a bandwidth h. The Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964) of $\pi(\mathbf{x})$ is given by

$$\tilde{\pi}(\mathbf{x}) = \frac{\sum_{j=1}^N K_h(\mathbf{x} - \mathbf{x}_j) R_j}{\sum_{j=1}^N K_h(\mathbf{x} - \mathbf{x}_j)}. \tag{4.6}$$

A kernel estimator in the form of $\tilde{\pi}(\mathbf{x})$ given in (4.6) usually has no practical values since we do not have complete auxiliary information for the finite population. It turns out that for the estimation of propensity scores the numerator in (4.6) only requires observations from the non-probability sample due to the binary variable R_j, and the denominator is a population total and can be estimated by using the reference probability sample. The nonparametric kernel regression estimator of the propensity scores is given by (Yuan, Li and Wu, 2022)

$$\hat{\pi}_i^A = \hat{\pi}(\mathbf{x}_i) = \frac{\sum_{j \in S_A} K_h(\mathbf{x}_i - \mathbf{x}_j)}{\sum_{j \in S_B} d_j^B K_h(\mathbf{x}_i - \mathbf{x}_j)}, \quad i \in S_A. \tag{4.7}$$

The estimator $\hat{\pi}_i^A$ given in (4.7) is consistent under the joint qp framework and the q-model for the propensity scores is very flexible due to the nonparametric assumption on $\pi(\mathbf{x})$. The estimated propensity scores are easy to compute when the dimension of $\mathbf{x}$ is not too high. Issues with high dimensional $\mathbf{x}$ and the choices of the kernel $K_h(\cdot)$ and the bandwidth h remain as in general applications of kernel-based estimation methods. Simulation results reported by Yuan et al. (2022) show that the kernel estimation method provides robust results for the propensity scores using the normal kernel and popular choices for the bandwidth.

Chu and Beaumont (2019) considered regression-tree based methods for estimating the propensity scores. Their proposed TrIPW method is a variant of the CART algorithm (Breiman, Friedman, Olshen and Stone, 1984) and uses data from the combined sample of the non-probability sample and the reference probability sample. The method aims to construct a classification tree with the terminal nodes of the final tree treated as homogeneous groups in terms of the propensity scores. The estimator of μ_y is constructed based on the final tree and post-stratification. Section 5 contains further details on poststratified estimators.

Statistical learning techniques such as classification and regression trees and random forests have been developed primarily for the purpose of prediction. Their use for estimating the propensity scores of non-probability samples requires further research. It is not a desirable approach to naively apply the methods over the pooled sample S_{AB} without theoretical justifications on the consistency of the final estimators. Further research towards this direction should be encouraged.

4.2 Inverse probability weighting

Let $\hat{\pi}_i^A$ be an estimate of $\pi_i^A = P\left(i \in S_A \mid \mathbf{x}_i\right)$ under a chosen method for the estimation of the propensity scores. Two versions of the inverse probability weighted (IPW) estimator of μ_y are constructed as

$$\hat{\mu}_{\text{IPW1}} = \frac{1}{N}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A} \quad \text{and} \quad \hat{\mu}_{\text{IPW2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i}{\hat{\pi}_i^A}, \tag{4.8}$$

where N is the population size and $\hat{N}^A = \sum_{i\in S_A}(\hat{\pi}_i^A)^{-1}$ is the estimated population size. The estimator $\hat{\mu}_{\text{IPW1}}$ is a version of the Horvitz-Thompson estimator and $\hat{\mu}_{\text{IPW2}}$ corresponds to the Hájek estimator as discussed in design-based estimation theory. There are ample evidences from both theoretical justifications and practical observations that the Hájek estimator $\hat{\mu}_{\text{IPW2}}$ performs better than the Horvitz-Thompson estimator and should be used in practice even if the population size N is known.

The validity of the IPW estimators $\hat{\mu}_{\text{IPW1}}$ and $\hat{\mu}_{\text{IPW2}}$ depends on the validity of the estimated propensity scores. Under the assumptions A1 and A2 and the parametric model $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$, the consistency of $\hat{\mu}_{\text{IPW1}}$ follows a standard two-step argument. Let $\tilde{\mu}_{\text{IPW}} = N^{-1}\sum_{i\in S_A} y_i/\pi_i^A$, which is not a computable estimator but an analytic tool useful for asymptotic purposes. It follows that $E_q\left(\tilde{\mu}_{\text{IPW}}\right) = \mu_y$ and the order $V_q\left(\tilde{\mu}_{\text{IPW}}\right) = O\left(n_A^{-1}\right)$ holds under the condition that $n_A\pi_i^A/N$ is bounded away from zero. As a consequence,

we have $\tilde{\mu}_{\text{IPW}} \to \mu_y$ in probability as $n_A \to \infty$. Under the correctly specified model $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}_0)$ for the propensity scores, the typical root-n order $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0 = O_p\left(n_A^{-1/2}\right)$ holds for commonly encountered scenarios. We can show by treating $\hat{\mu}_{\text{IPW1}}$ as a function of $\hat{\boldsymbol{\alpha}}$ and using a Taylor series expansion that $\hat{\mu}_{\text{IPW1}} = \tilde{\mu}_{\text{IPW}} + O_p\left(n_A^{-1/2}\right)$ under some mild finite moment conditions. The consistency of $\hat{\mu}_{\text{IPW2}}$ can be established using standard arguments for a ratio estimator (Section 5.3, Wu and Thompson, 2020) where $N^{-1}\sum_{i \in S_A}\left(\pi_i^A\right)^{-1} = 1 + o_p(1)$.

4.3 Doubly robust estimation

The dependence of the IPW estimator on the validity of the assumed propensity score model is viewed as a weakness of the method. The issue is not unique to the IPW estimators and is faced by many other approaches involving an assumed statistical model. Robust estimation procedures which provide certain degrees of protection against model misspecifications have been pursued by researchers, and the so-called doubly robust estimators have been a successful story since the work of Robins, Rotnitzky, and Zhao (1994).

The doubly robust (DR) estimator of μ_y is constructed using both the propensity score model q and the outcome regression model ξ. The DR estimator with the given propensity scores π_i^A, $i \in S_A$ and the mean responses $m_i = E_\xi\left(y_i \mid \mathbf{x}_i\right)$, $i = 1, 2, \ldots, N$ has the following general form,

$$\tilde{\mu}_{\text{DR}} = \frac{1}{N}\sum_{i \in S_A}\frac{y_i - m_i}{\pi_i^A} + \frac{1}{N}\sum_{i=1}^{N} m_i. \tag{4.9}$$

The second term on the right hand side of (4.9) is the model-based prediction of μ_y. The first term is a propensity score based adjustment using the errors $\varepsilon_i = y_i - m_i$ from the outcome regression model. The magnitude of the adjustment term is negatively correlated to the "goodness-of-fit" of the outcome regression model. It can be shown that $\tilde{\mu}_{\text{DR}}$ is an exactly unbiased estimator of μ_y if one of the two models q and ξ is correctly specified and hence it is doubly robust. The estimator $\tilde{\mu}_{\text{DR}}$ has an identical structure to the generalized difference estimator of Wu and Sitter (2001). It is important to note that the double robustness property of $\tilde{\mu}_{\text{DR}}$ does not require the knowledge of which of the two models being correctly specified. It is also apparent that the estimator $\tilde{\mu}_{\text{DR}}$ given in (4.9) is not computable in practical applications.

Let $\hat{\pi}_i^A$ and $\hat{m}_i$ be respectively the estimators of π_i^A and m_i under the assumed models q and ξ. Under the two-sample setting described in Section 2, the two DR estimators of μ_y proposed by Chen et al. (2020) are given by

$$\hat{\mu}_{\text{DR1}} = \frac{1}{N}\sum_{i \in S_A}\frac{y_i - \hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{i \in S_B} d_i^B \hat{m}_i \tag{4.10}$$

and

$$\hat{\mu}_{\text{DR2}} = \frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B}\sum_{i\in S_B} d_i^B\hat{m}_i, \tag{4.11}$$

where d_i^B are the design weights for the probability sample S_B, $\hat{N}^A=\sum_{i\in S_A}\left(\hat{\pi}_i^A\right)^{-1}$ and $\hat{N}^B=\sum_{i\in S_B} d_i^B$. The estimator $\hat{\mu}_{\text{DR2}}$ using the estimated population size has better performance in terms of bias and mean squared error and should be used in practice.

The probability survey design p is an integral part of the theoretical framework for assessing the two estimators $\hat{\mu}_{\text{DR1}}$ and $\hat{\mu}_{\text{DR2}}$. It is assumed that S_A and S_B are selected independently, which implies that $E_p\left(\sum_{i\in S_B} d_i^B\hat{m}_i\right)=\sum_{i=1}^N\hat{m}_i$. Consistency of the estimators $\hat{\mu}_{\text{DR1}}$ and $\hat{\mu}_{\text{DR2}}$ can be established under either the qp or the ξp framework. It should be noted that even if the non-probability sample S_A is a simple random sample with $\pi_i^A=n_A/N$, the doubly robust estimator in the form of (4.9) does not reduce to the model-based prediction estimator $\hat{\mu}_{y2}$ given in (3.3).

4.4 The pseudo empirical likelihood approach

The pseudo empirical likelihood (PEL) methods for probability survey samples have been under development over the past two decades. Two early papers on the topic are Chen and Sitter (1999) on point estimation incorporating auxiliary information and Wu and Rao (2006) on PEL ratio confidence intervals. The PEL approaches are further used for multiple frame surveys (Rao and Wu, 2010a) and Bayesian inferences with survey data (Rao and Wu, 2010b; Zhao, Ghosh, Rao and Wu, 2020b). Using the PEL methods for general inferential problems with complex surveys has been studied in two recent papers (Zhao and Wu, 2019; Zhao, Rao and Wu, 2020a).

Chen, Li, Rao and Wu (2022) showed that the PEL provides an attractive alternative approach to inference with non-probability survey samples. Let $\hat{\pi}_i^A$, $i\in S_A$ be the estimated propensity scores under an assumed parametric or non-parametric model, q. The PEL function for the non-probability survey sample S_A is defined as

$$\ell_{\text{PEL}}(\mathbf{p}) = n_A\sum_{i\in S_A}\tilde{d}_i^A\log\left(p_i\right), \tag{4.12}$$

where $\mathbf{p}=(p_1,\ldots,p_{n_A})$ is a discrete probability measure over the n_A selected units in S_A, $\tilde{d}_i^A=(\hat{\pi}_i^A)^{-1}/\hat{N}^A$ and $\hat{N}^A=\sum_{j\in S_A}(\hat{\pi}_j^A)^{-1}$ which is defined earlier in Section 4. Without using any additional information, maximizing $\ell_{\text{PEL}}(\mathbf{p})$ under the normalization constraint

$$\sum_{i\in S_A} p_i = 1 \tag{4.13}$$

leads to $\hat{p}_i=\tilde{d}_i^A$, $i\in S_A$. The maximum PEL estimator of μ_y is given by $\hat{\mu}_{\text{PEL}}=\sum_{i\in S_A}\hat{p}_i y_i$, which is identical to the IPW estimator $\hat{\mu}_{\text{IPW2}}$ given in (4.8).

The PEL approach to non-probability survey samples provides flexibilities in combining information through additional constraints and constructing confidence intervals and conducting hypothesis tests using the PEL ratio statistic. The maximum PEL estimator $\hat{\mu}_{\text{PEL}}=\sum_{i\in S_A}\hat{p}_i y_i$ is doubly robust if $\left(\hat{p}_1,\ldots,\hat{p}_{n_A}\right)$ is

the maximizer of $\ell_{\text{PEL}}(\mathbf{p})$ under both the normalization constraint and the model-calibration constraint given by

$$\sum_{i \in S_A} p_i \hat{m}_i = \bar{m}^B, \tag{4.14}$$

where $\bar{m}^B = \left(\hat{N}^B\right)^{-1} \sum_{i \in S_B} d_i^B \hat{m}_i$ is computed using the fitted values $\hat{m}_i$, $i \in S_B$ from an assumed outcome regression model, ξ. The equation (4.14) is a modified version of the original model-calibration constraint of Wu and Sitter (2001) using the probability sample S_B. Chen et al. (2022) contain further details on the asymptotic distributions of the PEL ratio statistic and simulation studies on the performances of PEL ratio confidence intervals on a finite population proportion.

5. Quota surveys and poststratification

Quota surveys are one of the oldest non-probability survey sampling methods which are still used in practice in present days. For a pre-specified overall sample size n_A, quotas of sample sizes are set for subpopulations which are defined by demographic variables and social-economic status indicators or other characteristic variables suitable for units of the target population. Data collection processes continue until quotas for each of the subpopulations are filled. Units from the population are typically approached using whatever convenient ways available and there are little or no controls on how units are selected for the final sample other than the pre-specified quotas.

The theory of the IPW estimators for non-probability survey samples provides an opportunity to examine scenarios where quota surveys may succeed or fail. For the convenience of notation without loss of generality, let S_A be the quota survey sample and $\mathbf{x}$ be the set of categorical variables used for defining the subpopulations and setting the quotas. The overall sample can be partitioned into $S_A = S_{A1} \cup \cdots \cup S_{AK}$ corresponding to the cross-classification of sampled units using the combinations of levels of the $\mathbf{x}$ variables. For instance, if $\mathbf{x} = (x_1, x_2)'$ with x_1 having two levels and x_2 having three levels, we have a total of $K = 2 \times 3 = 6$ subpopulations defined by $\mathbf{x}$. Let n_k be the pre-specified size of S_{Ak} and N_k be the size of the corresponding subpopulation. Under the assumption A1, the propensity scores $\pi_i^A = \pi(\mathbf{x}_i)$ become a constant for units in the same subpopulation and are given by $\pi_i^A = n_k / N_k$ for the k^{th} subpopulation. The IPW estimator $\hat{\mu}_{\text{IPW2}}$ given in (4.8) reduces to

$$\hat{\mu}_{\text{IPW2}} = \frac{1}{\hat{N}^A} \sum_{k=1}^{K} \sum_{i \in S_{Ak}} \frac{y_i}{\hat{\pi}_i^A} = \sum_{k=1}^{K} \hat{W}_k \bar{y}_k, \tag{5.1}$$

where $\bar{y}_k = n_k^{-1} \sum_{i \in S_{Ak}} y_i$, $\hat{W}_k = \hat{N}_k / \hat{N}^A$, $\hat{N}_k$ is the size of the k^{th} subpopulation obtained or estimated from external sources, and $\hat{N}^A = \sum_{k=1}^{K} \hat{N}_k$. Under the current setting with the availability of a reference probability sample S_B, we form the same partition as cross-classified by levels of $\mathbf{x}$ and obtain $S_B = S_{B1} \cup \cdots \cup S_{BK}$. We can then use $\hat{N}_k = \sum_{i \in S_{Bk}} d_i^B$.

The estimator given in (5.1) is the standard poststratified estimator of μ_y. It requires the information on the "stratum weights" $\hat{W}_k,\ k=1,\ldots,K,$ which is not available from the sample data itself. Quota surveys, combined with the use of the poststratified estimator, can be successful in producing valid population estimates for the study variable y if the following conditions hold:

(i) The categorical variables $\mathbf{x}$ used in defining the subpopulations and setting the quotas provide characterizations of the participation behavior of the units for voluntary surveys.

(ii) The inclusion of units in the survey is relatively random within each subpopulation and no specific groups are intentionally excluded from the survey.

(iii) The information on the stratum weights corresponding to the cross-classifications in setting the quotas can be reliably obtained from external sources.

(iv) The hardcore nonrespondents in the population who never take any voluntary surveys possess similar features to respondents in terms of the study variable y.

The IPW estimators $\hat{\mu}_{\text{IPW1}}$ and $\hat{\mu}_{\text{IPW2}}$ given in (4.8) may be sensitive to small values of estimated propensity scores. The poststratified estimator in the form of (5.1) serves as a robust alternative under general scenarios where the dimension of $\mathbf{x}$ is not low and/or some components of $\mathbf{x}$ are continuous. The K strata are formed based on homogeneous groups in terms of the propensity scores. Suppose that $\hat{\pi}_i^A=\pi(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}),\ i\in S_A$ are computed based on a parametric model, q. Suppose also that $n_A=m_AK$ with the chosen K where m_A is an integer. Let $\hat{\pi}_{(1)}^A\le\cdots\le\hat{\pi}_{(n_A)}^A$ be the estimated propensity scores in ascending order. Let S_{A1} be the set of the first m_A units in the sequence, S_{A2} be the second m_A units in the sequence, and so on. The poststratified estimator of μ_y is computed as $\hat{\mu}_{\text{PST}}=\sum_{k=1}^K\hat{W}_k\bar{y}_k,$ which has the same form of the estimator given in (5.1). The estimates of the stratum weights, $\hat{W}_k,\ k=1,2,\ldots,K$ can be obtained by using the reference probability sample S_B as follows. Let $b_k=\max\left\{\hat{\pi}_i^A\colon i\in S_{Ak}\right\},$ $k=1,2,\ldots,K-1.$ Let $b_0=0$ and $b_K=1.$

(a) Compute $\hat{\pi}_i=\pi(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}),\ i\in S_B.$

(b) Define $S_{Bk}=\left\{i\,\middle|\, i\in S_B, b_{k-1}<\hat{\pi}_i\le b_k\right\},\ k=1,2,\ldots,K.$

(c) Calculate $\hat{N}_k=\sum_{i\in S_{Bk}}d_i^B,\ k=1,2,\ldots,K.$

It is apparent that $S_B=S_{B1}\cup\cdots\cup S_{BK}$ and $\sum_{k=1}^K\hat{N}_k=\hat{N}^B=\sum_{i\in S_B}d_i^B.$ The estimated stratum weights are given by $\hat{W}_k=\hat{N}_k/\hat{N}^B.$

The choice of K needs to reflect the balance between homogeneity of the units within each post-stratum (in terms of the propensity scores) and the stability of the poststratified estimator (in terms of the stratum sample sizes). When the sample size n_A is small or moderate, a small number such as $K=5$ should be used. For scenarios where n_A is large, a larger K should be used such that units within the same poststratified sample S_{Ak} have similar estimated propensity scores. A practical guidance for the choice of K is to ensure that $m_A\ge30$ for the poststratified samples. For those who are old enough, do you remember the good old days when "the sample size is large" means "$n\ge30$"?

6. Variance estimation

Variance estimation under the two sample S_A and S_B setup involves at least two different sources of variation. The probability sampling design for the reference sample S_B remains one of the sources regardless of the approaches used for non-probability survey samples. Estimation of the variance component due to the use of S_B requires either suitable variance approximation formulas or replication weights as part of the dataset from the reference probability sample. Our discussion in this section assumes that a design-based variance estimator for the survey weighted point estimator based on S_B is available.

6.1 Variance estimation for mass imputation estimators

Variance estimation for the model-based prediction estimator $\hat{\mu}_y$ involves first deriving the asymptotic variance formula for $\text{Var}\left(\hat{\mu}_y - \mu_y\right)$ under the assumed outcome regression model or the imputation model ξ and the probability sampling design p, and then using plug-in estimators for various unknown population quantities.

The mass imputation estimator $\hat{\mu}_{y\text{MI}} = \hat{N}_B^{-1}\sum_{i\in S_B} d_i^B y_i^*$ given in (3.5) is a special type of model-based prediction estimator, where the model ξ refers to the one used for imputation and is not necessarily the same as the outcome regression model. The imputation method plays a key role in deriving the asymptotic variance formula, and the variance estimator needs to be constructed accordingly. Noting that $\hat{\mu}_{y\text{MI}}$ is a Hájek type estimator due to the use of the estimated population size $\hat{N}_B$, derivations of the asymptotic variance formula start with putting the true value N in first and then dealing with $\hat{\mu}_{y\text{MI}}$ as a ratio estimator. Kim et al. (2021) considered variance estimation for $\hat{\mu}_y = N^{-1}\sum_{i\in S_B} d_i^B y_i^*$, where $y_i^* = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$ is the imputed value for y_i based on the semiparametric model (3.1). The asymptotic variance formula is developed in two steps. First, a linearized version of $\hat{\mu}_y$ is obtained by using a Taylor series expansion at $\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$. Second, two variance components are derived for $\text{Var}\left(\hat{\mu}_y - \mu_y\right)$ based on the linearized version using the semiparametric model (3.1) and the sampling design for S_B. The process is tedious, which is the case for most model-based variance estimation methods. A bootstrap variance estimator turns out to be more attractive for practical applications. See Kim et al. (2021) for further details.

6.2 Variance estimation for IPW estimators

The commonly used IPW estimator $\hat{\mu}_{\text{IPW2}}$ given in (4.8) is valid under the assumed model q for the propensity scores. An explicit asymptotic variance formula for $\hat{\mu}_{\text{IPW2}}$ can be derived under the joint qp-framework when the propensity scores are estimated using the pseudo maximum likelihood method or an estimating equation based method as discussed in Section 4.1. The theoretical tool is the sandwich-type variance formula for point estimators defined as the solution to a combined system of estimating equations for both μ_y and $\boldsymbol{\alpha}_0$.

Consider the parametric form $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ for the propensity scores, where the model parameters $\boldsymbol{\alpha}$ are estimated through the estimating equations (4.4) with user-specified functions $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha})$. The first major step in deriving the asymptotic variance formula for $\hat{\mu}_{\text{IPW2}}$ is to write down the system of joint estimating equations for both μ_y and $\boldsymbol{\alpha}_0$. Let $\boldsymbol{\eta} = (\mu, \boldsymbol{\alpha}')'$ be the vector of the combined parameters. The estimator $\hat{\boldsymbol{\eta}} = (\hat{\mu}_{\text{IPW2}}, \boldsymbol{\alpha}')'$ is the solution to the system of joint estimating equations $\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = \mathbf{0}$, where

$$\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = \begin{pmatrix} N^{-1}\sum_{i=1}^N R_i(y_i - \mu)/\pi_i^A \\ N^{-1}\sum_{i=1}^N R_i \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - N^{-1}\sum_{i \in S_B} d_i^B \pi_i^A \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}. \tag{6.1}$$

The factor N^{-1} is redundant but useful in facilitating asymptotic orders. The estimating functions defined by (6.1) are unbiased under the joint qp-framework, i.e., $E_{qp}\{\boldsymbol{\Phi}(\boldsymbol{\eta}_0)\} = \mathbf{0}$, where $\boldsymbol{\eta}_0 = (\mu_y, \boldsymbol{\alpha}_0')'$. There are two major consequences from the unbiasedness of the estimating equations system. First, consistency of the estimator $\hat{\boldsymbol{\eta}}$ can be argued using the theory of general estimating functions similar to those presented in Section 3.2 of Tsiatis (2006). Second, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\eta}}$, denoted as $\text{AV}(\hat{\boldsymbol{\eta}})$, has the standard sandwich form and is given by

$$\text{AV}(\hat{\boldsymbol{\eta}}) = \left[E\{\dot{\boldsymbol{\phi}}_n(\boldsymbol{\eta}_0)\}\right]^{-1} \text{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\} \left[E\{\dot{\boldsymbol{\phi}}_n(\boldsymbol{\eta}_0)\}'\right]^{-1},$$

where $\dot{\boldsymbol{\phi}}_n(\boldsymbol{\eta}) = \partial \boldsymbol{\Phi}_n(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$, which depends on the forms of $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ and $\mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha})$. The term $\text{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\}$ consists of two components, one due to the propensity score model q and the other from the probability sampling design for S_B. More specifically, we have $\text{Var}\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\} = V_q(\mathbf{A}_1) + V_p(\mathbf{A}_2)$, where $V_q(\cdot)$ denotes the variance under the propensity score model q and $V_p(\cdot)$ represents the design-based variance under the probability sampling design p, and

$$\mathbf{A}_1 = \frac{1}{N}\sum_{i=1}^N R_i \begin{pmatrix} (y_i - \mu)/\pi_i^A \\ \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}, \quad \mathbf{A}_2 = \frac{1}{N}\sum_{i \in S_B} d_i^B \begin{pmatrix} 0 \\ \pi_i^A \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) \end{pmatrix}.$$

The analytic expression for $V_q(\mathbf{A}_1)$ follows immediately from $V_q(R_i) = \pi_i^A(1 - \pi_i^A)$ and the independence among $R_1, \ldots, R_N$. The design-based variance component $V_p(\mathbf{A}_2)$ requires additional information on the survey design for S_B or a suitable variance approximation formula with the given design.

The asymptotic variance formula for the IPW estimator $\hat{\mu}_{\text{IPW2}}$ is the first diagonal element of the matrix $\text{AV}(\hat{\boldsymbol{\eta}})$. The final variance estimator for $\hat{\mu}_{\text{IPW2}}$ can then be obtained by replacing various population quantities with sample-based moment estimators. Chen et al. (2020) presented the variance estimator with explicit expressions when $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ are modelled by the logistic regression and the $\hat{\boldsymbol{\alpha}}$ is obtained by the pseudo maximum likelihood method.

6.3 Variance estimation for doubly robust estimators

It turns out that variance estimation for the doubly robust estimator is a challenging problem. While double robustness is a desirable property for point estimation, it creates a dilemma for variance estimation.

The estimator $\hat{\mu}_{\text{DR2}}$ given in (4.11) is consistent if either the propensity score model q or the outcome regression model ξ is correctly specified. There is no need to know which model is correctly specified, which is the most crucial part behind double robustness. This ambiguous feature, however, becomes a problem for variance estimation. The asymptotic variance formula under the model q is usually different from the one under the model ξ, and consequently, it is difficult to construct a consistent variance estimator with unknown scenarios on model specifications.

There have been several strategies proposed in the literature on variance estimation for the doubly robust estimators. A naive approach is to use the variance estimator derived under the assumed propensity score model q and take the risk that such a variance estimator might have non-negligible biases under the outcome regression model. One good news is that, under the propensity score model, the estimation of the parameters $\boldsymbol{\beta}$ for the outcome regression model has no impact asymptotically on the variance of doubly robust estimators. This can be seen by using $\hat{\mu}_{\text{DR1}}$ of (4.10) as an example. Let $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}\right)$, where $\hat{\boldsymbol{\beta}}$ is obtained based on the working model (3.1) which is not necessarily correct. Let $\boldsymbol{\beta}^*$ be the probability limit of $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p\left(n_A^{-1/2}\right)$ regardless of the true outcome regression model (White, 1982). Let $m_i^* = m\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$ and $\mathbf{a}(\mathbf{x}, \boldsymbol{\beta}) = \partial m(\mathbf{x}, \boldsymbol{\beta}) / \partial \boldsymbol{\beta}$. It can be seen that

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \; + \; \left\{\mathbf{B}\left(\boldsymbol{\beta}^*\right)\right\}' \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right) \; + \; o_p\left(n_A^{-1/2}\right),$$

where

$$\mathbf{B}\left(\boldsymbol{\beta}^*\right) \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right) \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)}{\hat{\pi}_i^A}. \tag{6.2}$$

Since the two terms on the right hand side of (6.2) are both consistent estimators of $N^{-1}\sum_{i=1}^N \mathbf{a}\left(\mathbf{x}_i, \boldsymbol{\beta}^*\right)$, we conclude that $\mathbf{B}\left(\boldsymbol{\beta}^*\right) = o_p(1)$ and

$$\frac{1}{N}\sum_{i\in S_B} d_i^B \hat{m}_i \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{\hat{m}_i}{\hat{\pi}_i^A} \; = \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; - \; \frac{1}{N}\sum_{i\in S_A} \frac{m_i^*}{\hat{\pi}_i^A} \; + \; o_p\left(n_A^{-1/2}\right).$$

It follows that

$$\hat{\mu}_{\text{DR1}} \; = \; \frac{1}{N}\sum_{i\in S_A} \frac{y_i - m_i^*}{\hat{\pi}_i^A} \; + \; \frac{1}{N}\sum_{i\in S_B} d_i^B m_i^* \; + \; o_p\left(n_A^{-1/2}\right).$$

The same arguments apply to $\hat{\mu}_{\text{DR2}}$. We can treat $\hat{\boldsymbol{\beta}}$ as if it is fixed in deriving the asymptotic variance for $\hat{\mu}_{\text{DR1}}$ and $\hat{\mu}_{\text{DR2}}$ under the assumed propensity score model. The techniques described in Section 6.2 can be directly used where the first estimating function in (6.1) is replaced by the one for defining $\hat{\mu}_{\text{DR1}}$ or $\hat{\mu}_{\text{DR2}}$. See Theorem 2 of Chen et al. (2020) for further details. The variance estimator derived under the propensity score model, however, is generally biased under the outcome regression model.

Chen et al. (2020) also described a technique using the original idea presented in Kim and Haziza (2014) for the construction of the so-called doubly robust variance estimator. The technique is a delicate one with some theoretical attractiveness but has various issues for practical applications. We use $\hat{\mu}_{\text{DR1}}$ as an example to illustrate the steps for the construction of the doubly robust variance estimator. Let

$$\hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} R_i \frac{y_i - m(\mathbf{x}_i, \boldsymbol{\beta})}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} + \frac{1}{N}\sum_{i \in S_B} d_i^B m(\mathbf{x}_i, \boldsymbol{\beta}).$$

It follows that $\hat{\mu}_{\text{DR1}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ if $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are from the original estimation methods. The first step is to modify the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are obtained as solutions to

$$\frac{\partial \hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\alpha}} = \mathbf{0} \quad \text{and} \quad \frac{\partial \hat{\mu}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{6.3}$$

Under the logistic regression model q where $\text{logit}\{\pi(\mathbf{x}_i, \boldsymbol{\alpha})\} = \mathbf{x}_i'\boldsymbol{\alpha}$ and the linear regression model ξ where $m(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}$, the equation system (6.3) becomes

$$\frac{1}{N}\sum_{i=1}^{N} R_i \left\{\frac{1}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - 1\right\}(y_i - \mathbf{x}_i'\boldsymbol{\beta})\,\mathbf{x}_i = \mathbf{0}, \tag{6.4}$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i \mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \frac{1}{N}\sum_{i \in S_B} d_i^B \mathbf{x}_i = \mathbf{0}. \tag{6.5}$$

The estimating equations in (6.5) are unbiased under the joint qp-framework. They are identical to (4.5) discussed in Section 4.1.2. The estimating equations in (6.4) are also unbiased under the outcome regression model, but they are different from the quasi score equations given in (3.2). The estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ obtained as solutions to (6.4) and (6.5) are less stable than those from standard methods. In addition, the equations system (6.4) and (6.5) will not have a solution if $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not of the same dimension, since the number of equations in (6.4) is decided by the dimension of $\boldsymbol{\alpha}$ and the number of equations in (6.5) is the same as the dimension of $\boldsymbol{\beta}$. The final estimator $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ also suffers from efficiency losses when $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by solving (6.4) and (6.5).

The reason behind the use of the equations system (6.3) is purely technical. It can be shown through a first order Taylor series expansion that the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ obtained from (6.3) have no impact asymptotically on the variance of $\hat{\mu}_{\text{DR}} = \hat{\mu}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$. This technical maneuver enables that simple explicit expressions for the variance $V_{qp}(\hat{\mu}_{\text{DR}})$ under the qp framework and for the prediction variance $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ under the ξp framework can easily be obtained. Construction of the doubly robust variance estimator for $\hat{\mu}_{\text{DR}}$ starts with the plug-in estimator for $V_{qp}(\hat{\mu}_{\text{DR}})$ under the propensity scores model q. A bias-correction term is then added to obtain a valid estimator for $V_{\xi p}(\hat{\mu}_{\text{DR}} - \mu_y)$ under the outcome regression model ξ. The happy ending of the story is that the bias-correction term has the analytic form $N^{-2}\sum_{i=1}^{N}(R_i/\pi_i^A - 1)\sigma_i^2$ where $\sigma_i^2 = E_\xi(y_i \mid \mathbf{x}_i)$, which is negligible under the propensity

score model q. The bias-corrected variance estimator is valid under either the propensity score model or the outcome regression model.

A doubly robust variance estimator for the commonly used $\hat{\mu}_{\text{DR2}}$ is not available in the literature. A practical solution is to use bootstrap methods. Chen et al. (2022) demonstrated that standard with-replacement bootstrap procedures applied separately to S_A and S_B provide doubly robust confidence intervals using the pseudo empirical likelihood approach to non-probability survey samples when the reference sample is selected by single stage unequal probability sampling designs. Complications will arise when the probability sample S_B uses stratified multi-stage sampling methods, a known challenge for variance estimation with complex surveys. Construction of doubly robust variance estimators for the doubly robust estimator $\hat{\mu}_{\text{DR2}}$ under general settings deserves efforts in future research.

7. Assumptions revisited

Our discussions on estimation procedures for non-probability survey samples are under the assumptions A1-A4 and the focuses are on the validity and efficiency of estimators for the finite population mean under three inferential frameworks. The theoretical results on model-based prediction, inverse probability weighting and doubly robust estimation have been rigorously established under those assumptions. It seems that researchers are triumphant in dealing with the emerging area of non-probability data sources. However, as pointed out by the 2021 ASA President Robert Santos in his opinion article entitled "Using Our Superpowers to Contribute to the Public Good" (Amstat News, May 2021), "*Our superpowers are only as good as their underlying assumptions, assumptions that are all too often embraced with aplomb, yet cannot be proven*." How to check assumptions A1-A4 in practical applications of the methods is a question that can never be fully answered, and yet there are steps to follow to boost the confidence in using the theoretical results. It is also important to understand the potential consequences when certain assumptions become seriously questionable.

7.1 Assumption A1

Assumption A1 states that $\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i\right)$. It is the most crucial assumption for the validity of the pseudo maximum likelihood estimator of Chen et al. (2020) and the nonparametric kernel smoothing estimator presented in Section 4.1.3 for the propensity scores, although all other assumptions are also involved. It is equivalent to the missing at random (MAR) assumption in the missing data literature. It is well understood that the MAR assumption cannot be tested using the sample data itself. The same statement holds for assumption A1 with non-probability survey samples.

In a nutshell, assumption A1 indicates that the auxiliary variables $\mathbf{x}$ included in the non-probability sample fully characterize the participation behaviour or the sample inclusion mechanism for units in the population. Sufficient attention should be given at the study design stage before data collection, if such a stage exists, to investigate potential factors and features of units which might be related to participation

and sample inclusion. For human populations, the factors and features may include demographical variables, social and economic indicators, and geographical variables.

Assumption A1 leads to the conclusion that the conditional distribution of y given $\mathbf{x}$ for units in the non-probability sample is the same as the conditional distribution of y given $\mathbf{x}$ for units in the target population. It implies that the auxiliary variables $\mathbf{x}$ should include relevant predictors for the study variable y. With the given datasets S_A and S_B, sensitivity analysis through comparisons of marginal distributions and conditional models can be helpful in building confidence on assumption A1. For variables which are available in both S_A and S_B, one can compare the empirical distribution functions (or moments) from S_A to the survey weighted empirical distribution functions (or moments) from S_B. Marked differences between the two indicate that S_A is a non-probability sample with unequal propensity scores. One possible sensitivity analysis on assumption A1 is to select a variable z which has certain similarities to y, and a set of auxiliary variables $\mathbf{u}$ with both z and $\mathbf{u}$ available from S_A and S_B. We fit a conditional model $z \mid \mathbf{u}$ using data from S_A and a survey weighted conditional model $z \mid \mathbf{u}$ using data from S_B. If $\mathbf{u}$ includes all the key auxiliary variables for assumption A1, we should see the two versions of fitted models to be similar to each other. Drastic differences between the two fitted models are a strong sign that either the z is itself an important auxiliary variable for assumption A1 or the assumption is questionable.

7.2 Assumption A2

A casual look at assumption A2 may have people believe that it should easily be satisfied in practice, since a similar assumption is widely used in missing data analysis and causal inference. It turns out that the assumption can be highly problematic, and for scenarios where the assumption fails to hold, the target population is different from the one assumed for the estimation methods. It is similar to the frame undercoverage and nonresponse problems which are discussed extensively in probability sampling.

Assumption A2 states that $\pi_i^A = P\left(R_i = 1 \mid \mathbf{x}_i, y_i\right) > 0$ for all i. It is equivalent to stating that every unit in the target population has a non-zero probability to be included in the non-probability sample. If the sample was taken by a probability sampling method, this would be the scenario where the sampling frame is complete and there are no hardcore nonrespondents. For most non-probability samples, the concept of "*sampling frame*" is often irrelevant or simply a convenient list, and the selection and inclusion of units for the sample may not have a structured process. In her presentation at the 2021 CANSSI-NISS Workshop, Mary Thompson pointed out that "*the statement that the sample inclusion indicator R is a random variable is itself an assumption*" for non-probability survey samples.

Let U be the set of N units for the target population. Let $U_0 = \left\{ i \mid i \in U \text{ and } \pi_i^A > 0 \right\}$. It is apparent that $U_0 \subset U$ and $U_0 \neq U$ when assumption A2 is violated. There are two typical scenarios in practice. The first can be termed as *stochastic undercoverage*, where the non-probability sample S_A is selected from U_0 and U_0 itself can be viewed as a random sample from U. For example, the contact list of an existing probability survey is used to approach units in the population for participation in the non-

probability sample. In this case U_0 consists of units from the probability sample. Another example is a volunteer survey where the target population consists of adults in a specific city/region but the participants are recruited from visitors to major shopping centers in the region over certain period of time. The subpopulation U_0 includes visitors to the chosen locations over the sampling period and it is reasonable to assume that U_0 is a random sample from the target population. Let $D_i = 1$ if $i \in U_0$ and $D_i = 0$ otherwise, $i = 1, 2, \ldots, N$. We have

$$P\left(R_i = 1 \middle| \mathbf{x}_i, y_i, D_i = 1\right) > 0 \quad \text{and} \quad P\left(R_i = 1 \middle| \mathbf{x}_i, y_i, D_i = 0\right) = 0$$

for $i = 1, 2, \ldots, N$. If the subpopulation U_0 is formed with an underlying stochastic mechanism such that $P\left(D_i = 1 \middle| \mathbf{x}_i, y_i\right) > 0$ for all $i \in U$, we have

$$\pi_i^A = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i\right) = P\left(R_i = 1 \middle| \mathbf{x}_i, y_i, D_i = 1\right) P\left(D_i = 1 \middle| \mathbf{x}_i, y_i\right) > 0$$

for $i = 1, 2, \ldots, N$. In other words, the assumption A2 is valid under the scenario of stochastic undercoverage for non-probability samples.

The second scenario is termed as *deterministic undercoverage* where units with certain features will never be included in the non-probability sample. Suppose that participation in the non-probability survey requires internet access and a valid email address, and 20% of the population have neither access to the internet nor an email address, we have an example where the 20% of the population have zero propensity scores. There is no simple fix to the inferential procedures developed under A2. Yilin Chen's PhD dissertation at University of Waterloo (Chen, 2020) contained one chapter dealing with some specific aspects of the scenario.

7.3 Assumption A3

Among all the assumptions, this one is less crucial to the validity of the proposed inferential procedures. Under assumption A3, the full likelihood function for the propensity scores is given in (4.1). For any parametric model on $\pi_i^A = \pi\left(\mathbf{x}_i, \boldsymbol{\alpha}\right)$, the quasi log-likelihood function $\ell^*(\boldsymbol{\alpha})$ given in (4.2) leads to the quasi score functions $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$, which remains unbiased even if assumption A3 is violated. There might be some efficiency loss without assumption A3 in estimating the model parameters $\boldsymbol{\alpha}$ but the estimation methods are still valid under the other three assumptions.

7.4 Assumption A4

It is not difficult to find an existing probability sample from the same target population. It might be very hard, however, to have a probability survey sample which contains the desirable auxiliary variables. Existing probability surveys are designed with specific aims and scientific objectives, and the auxiliary variables included in the survey are not necessarily relevant to the analysis of a particular non-probability survey sample. The ultimate goal for satisfying assumption A4 is to identify and gain access to an existing

probability survey sample with a rich collection of demographical variables, social and economic indicators, and geographical variables.

A rich-people's problem (when one has too much money) for assumption A4 may also occur in practice when two or more existing probability survey samples are available. How to combine all of them for more efficient analysis of non-probability survey samples is a research topic that deserves further attention. Some practical guidances on choosing one reference probability sample from available alternatives include following considerations.

(i) Check for availability of important auxiliary variables which are relevant to characterizing the participation behavor or having prediction power to the study variables in the non-probability sample;

(ii) Give first preference to the one with a larger set of variables that are common to the non-probability sample;

(iii) Assign second preference to the probability sample with a larger sample size;

(iv) And lastly, use the probability sample for which the mode of data collection is the same as the one for the non-probability sample.

It was shown by Chen et al. (2020) that two reference probability survey samples with the same set of common auxiliary variables tend to produce very similar IPW estimators but the one with a larger sample size leads to better mass imputation estimators.

8. Concluding remarks

In the early years of the 21st century, Web-based surveys started to become popular, which generated substantial amount of research interest on the topic (Tourangeau, Conrad and Couper, 2013). Issues and challenges faced by web-based and other non-probability survey samples led to the "Summary Report of the AAPOR Task Force on Non-probability Sampling" by Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013). Among other things, the report indicated that (i) unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling; (ii) making inferences for any probability or non-probability survey requires some reliance on modeling assumptions; and (iii) 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.

Survey sampling researchers have been answering the call with intensified explorations on statistical inference with non-probability survey samples. The current setting of two samples S_A and S_B, with the non-probability sample S_A having measurements on both the study variable y and auxiliary variables $\mathbf{x}$ and the probability sample S_B providing information on $\mathbf{x}$, was first considered by Rivers (2007) on sample matching using nearest neighbor imputation, which is the original idea leading to the mass

imputation method (Kim et al., 2021). The weighted logistic regression using the pooled sample for estimating the propensity scores proposed by Valliant and Dever (2011) was the first serious attempt on the topic, which serves as a motivation for the pseudo maximum likelihood method developed by Chen et al. (2020). Brick (2015) considered compositional model inference under the same setting. Elliott and Valliant (2017) provided informed discussions on inference for non-probability samples. Yang, Kim and Song (2020) addressed issues with high dimensional data in combining probability and non-probability survey samples.

Statistical inference with non-probability survey samples is part of the more general topic on combining data from multiple sources. The term "data integration" is frequently used under this context. Combining information from independent probability survey samples has been studied extensively in the survey literature; see, for instance, Wu (2004), Kim and Rao (2012) and references therein. Inferences with samples from multiple frame surveys are another topic which has been heavily investigated by survey statisticians; see Lohr and Rao (2006) and Rao and Wu (2010a) and references therein. In her recent Waksberg award invited paper, Lohr (2021) provided an overview on multiple-frame surveys and some fascinating discussions on using a multiple-frame structure to serve as an organizing principle for other data combination methods. With emerging new data sources and reshaped views on traditional data sources such as administrative records, data integration has become a very broad area that calls for continued research. Further discussions are provided by Lohr and Raghunathan (2017) on combining survey data with other data sources and by Thompson (2019) on combining new and traditional sources in population surveys. Kim and Tam (2021) and Yang, Kim and Hwang (2021) discussed data integration by combining big data and survey sample data for finite population inference. Yang and Kim (2020) contained a review on statistical data integration in survey sampling.

One of the essential messages that the current paper conveys is the concepts of *validity* and *efficiency* in analyzing non-probability survey samples. Validity refers to the consistency of point estimators and efficiency is measured by the asymptotic variance of the point estimator. Validity is of primary concern and efficiency pursuit is a secondary goal when valid alternative approaches are available. Discussions on validity and efficiency require a suitable inferential framework and rigorous developments of statistical procedures, which is another main message from this paper. Non-probability samples do not fit into the traditional design-based or model-based inferential framework for probability survey samples. Standard statistical concepts and inferential procedures, however, can be built into a suitable framework for valid and efficient inference with non-probability survey samples.

Non-probability samples may have a very large sample size. Large sample sizes are a double-edged sword: when the inferential procedures are valid, large sample sizes lead to more efficient inference; when the estimators are biased, large sample sizes make the bias even more pronounced. A non-probability survey sample with a 80% sampling fraction over the population does not necessarily provide better estimation results than a small probability sample (Meng, 2018).

The large sample sizes also make non-probability samples connected to the modern big data problems. The role of traditional statistical methods in the era of big data was convincingly argued by Richard Lockhart (2018): "*Huge new computing resources do not put an end to the need for careful modelling, for honest assessment of uncertainty, or for good experiment design. Classical statistical ideas continue to have a crucial role to play in keeping data analysis honest, efficient, and effective.*"

Jean-François Beaumont (2020) raised the question "Are probability surveys bound to disappear for the production of official statistics?" The short answer is that probability sampling methods and probability survey samples will remain as an important data collection tool for many fields, including official statistics, and design-based inference will play a crucial role for any evolving inferential framework. The current trend of using non-probability samples and data from other sources will continue. Valid and efficient statistical inference with non-probability samples requires auxiliary information from the target population. A few high quality national probability surveys with carefully designed survey variables can play a pivotal role in analysis of non-probability survey samples.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Statistical Sciences Institute. An early version of the paper was presented at the SSC 2021 Annual Meeting as the Special Presidential Invited Address by the Survey Methods Section of the SSC. The author thanks the Editor of *Survey Methodology*, Jean-François Beaumont, for the invitation and for organizing the discussions on the emerging topic of statistical inference with non-probability survey samples. Thanks are also due for the two anonymous reviewers who provided constructive comments on the initial submission which led to improvements of the paper.

References

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

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/pub/12-001-x/2020001/article/00001-eng.pdf.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, second edition. Wadsworth & Brooks/Cole Advanced Books & Software.

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

Chen, J., and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*,16, 113-131.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association*, 96, 260-269.

Chen, J., and Sitter, R.R. (1999). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, Y. (2020). *Statistical Analysis with Non-probability Survey Samples*, PhD Dissertation, Department of Statistics and Actuarial Science, University of Waterloo.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chen, Y., Li, P., Rao, J.N.K. and Wu, C. (2022). Pseudo empirical likelihood inference for non-probability survey samples. *The Canadian Journal of Statistics*, accepted.

Chu, K.C.K., and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. Proceedings of the Survey Methods Section of SSC.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics*, 31, 1208-1212.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Kim, J.K., and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24, 375-394.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99, 85-100.

Kim, J.K., and Tam, S. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89, 382-401.

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.

Liu, Z., and Valliant, R. (2021). Investigating an alternative for estimation from a nonprobability sample: Matching plus calibration. arXiv:2112.00855v1 [stat.ME]. Dec. 2021.

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46, 4-9.

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., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, second edition, New York: Chapman and Hall.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.

Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Applications*, 9, 141-142.

Neyman, J. (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, 558-606.

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

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*, second Edition. Hoboken, NJ: Wiley.

Rao, J.N.K., and Wu, C. (2010a). Pseudo empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

Rao, J.N.K., and Wu, C. (2010b). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society, Series B*, 72, 533-544.

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.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*,89, 846-866

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Tourangeau, R., Conrad, F.G. and Couper, M.P. (2013). *The Science of Web Surveys*, first edition. Oxford: Oxford University Press.

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

Thompson, M.E. (2019). Combining data from new and traditional sources in population surveys. *International Statistical Review*, 87, S79-89.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

Wang, L., Graubard, B.I., Katki, H.A. and Li, Y. (2020). Improving external validity of epidemiologic cohort analysis: A kernel weighting approach. *Journal of the Royal Statistical Society, Series A*, 183, 1293-1311.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

Watson, G.S. (1964). Smooth regression analysis. *Sankhyā A*, 26, 359-372.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-25.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, 32, 15-26.

Wu, C., and Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34, 359-375.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

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

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2021001/article/00004-eng.pdf.

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B*, 82, 445-465.

Yuan, M., Li, P. and Wu, C. (2022). Nonparametric estimation of propensity scores for non-probability survey samples. Working paper.

Zhao, P., and Wu, C. (2019). Some theoretical and practical aspects of empirical likelihood methods for complex surveys. *International Statistical Review*, 87, S239-256.

Zhao, P., Rao, J.N.K. and Wu, C. (2020a). Empirical likelihood methods for public-use survey data. *Electronic Journal of Statistics*, 14, 2484-2509.

Zhao, P., Ghosh, M., Rao, J.N.K. and Wu, C. (2020b). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society, Series B*, 82, 155-174.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 313-318
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical inference with non-probability survey samples" – Non-probability samples: An assessment and way forward

Michael A. Bailey[1]

Abstract

Non-probability surveys play an increasing role in survey research. Wu's essay ably brings together the many tools available when assuming the non-response is conditionally independent of the study variable. In this commentary, I explore how to integrate Wu's insights in a broader framework that encompasses the case in which non-response depends on the study variable, a case that is particularly dangerous in non-probabilistic polling.

Key Words: Survey sampling; Non-probability polls.

1. Introduction

Surveys are going through massive changes. Gone are the days of random digit dialing phone surveys producing reliably representative samples. Now hardly anyone answers the phone or even responds to emails. Pollsters have responded by coming up with a myriad of clever new ways to generate survey responses in this unwelcoming environment.

The most pervasive innovation is, without a doubt, the use of non-probability samples, often via the internet. While the implementation varies, the approach typically gathers contact information for a large number of people who are willing to respond and then involves selecting a subset from that pool for any given survey. These surveys have proven cost-effective and have often – if, perhaps, not always – produced serviceable results.

But are they believable? Most surveys do not have a ground truth against which to assess results; the lack of such information is, after all, the reason why someone is conducting the survey. Probability samples overcome this problem by relying on theory as the properties of such surveys are well understood. For non-probability samples, however, practice has vastly outpaced theory, meaning that the basis for believing the results is rather speculative.

Wu's paper therefore is a welcome contribution to our understanding of non-probability surveys. He focuses on the class of estimators that assume ignorable non-response and puts them in context relative to each other and identifies avenues for future work.

One important point made by Wu is that "there must be a more coherent framework and accompanying set of measures for evaluating their quality" (page 305). I heartily concur. In this commentary, I expand on this point in three ways. In Section 2 I explore how to do this within the scope of the research he

1. Michael A. Bailey, Georgetown University. E-mail: baileyma@georgetown.edu.

examines. In Section 3 I seek to expand the scope of such a framework, noting that the consequences of violations of key assumptions are so much more severe in a non-probability setting that we should build our framework to encompass violations of the key missing-at-random (MAR) assumption. In Section 4 I then explore what, if anything we can do about it. Finally, in Section 5 I provide a few concluding remarks.

2. Non-probability surveys when data is MAR

Wu grounds his analysis with a clear exposition of the four assumptions underlying the models he examines. The most important assumption is that data is MAR, meaning that given a set of covariates the study variable is independent of the decision to respond. (Although the nomenclature is standard in the literature, I cannot resist registering unease with the "missing at random" label. Of course, data is missing at random – something that is true even for MAR's opposite (and also inaptly named) missing-not-at-random (MNAR). I dream of a day when the nomenclature matches the definition, perhaps by replacing MAR with the term "conditional independence" would be a better name. However, I recognize how hard it is to change the accepted terms people use.)

Given these assumptions, Wu divides approaches into those that are model-based, inverse propensity weighting (IPW) based and double robust models. In the model-based approaches, we see the range of efforts to impute from the observed sample, including mass imputation that, broadly conceived, includes flexible sample-matching approaches that allow us to represent a larger population based on observed data points that are "close", variously defined. IPW builds on the same assumptions. Doubly robust estimators tend to be newer and attractive for their ability to give researchers two bites at the apple of relying on correct assumptions; Wu ably documents the headaches these models bring when we try to do inference with them, however.

While Wu has shown the differences in these approaches, it is useful to appreciate that he is fishing in one fairly specific corner of the pond. All models use similar information in similar ways: they all assume MAR and provide tools to model or impute the behavior of unobserved people as direct extrapolations from the observed data. If college graduates differ from non-college graduates and we have too many college graduates, all the MAR-based approaches will extrapolate to the general population directly from the data in the sample on two groups in proportion to these groups' presence in the target population.

My intuition is that the models considered by Wu are roughly equally useful – and also roughly vulnerable to violations of MAR. Or, are there contexts in which we expect the differences across the methods to be substantial? Answering this is not easy, of course, but I would be fascinated to learn Wu's perspective on where the main "action" is in non-probability samples and which of the models he considers would be best suited to accounting for such problems.

One possible focus would be on the flexibility across models. At this point, my intuition is that while these differences could be substantial in theory, in practice these differences are relatively modest. This is

especially true if an experienced researcher with domain knowledge specifies a parametric model with a deft touch – including the right interactions and so forth.

3. Non-probability surveys when data is not MAR

We should take very seriously Wu's call for more coherent framework for analyzing non-probability samples. And we should aim big here as a paradigm for non-probability samples is, essentially, a paradigm for the whole field given the importance and trajectory of non-probability samples.

As we think about formulating a framework for polling it is useful to recall George Box's famous aphorism: "Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad" (Box, 1976). The tiger in non-probability samples does not live between quota sampling and IPW models. The tiger can almost certainly be found instead in the MAR assumption. The violation of this assumption is the signature weakness of MAR and any framework for non-probability surveys should therefore start there.

The issue is that while MAR violations are a problem in probability sampling (arising due to non-response among the randomly contacted individuals), MAR violations are more serious in a non-probabilistic world. The idea is formalized in Meng (2018) who provides an identity for the error in a survey:

$$\bar{Y}_n - \bar{Y}_N = \underbrace{\rho_{R,Y}}_{\text{data quality}} \underbrace{\sqrt{\frac{N-n}{n}}}_{\text{data quantity}} \underbrace{\sigma_Y}_{\text{data difficulty}} . \tag{3.1}$$

The first term in the equation is $\rho_{R,Y}$, the correlation in the population between R and Y. This quantity can be taken to reflect quality of data with regard to sampling. The second term in the Meng equation, $\sqrt{N-n/n}$, relates to the size of the population (capital N) and the size of the sample (lower case n). The third term in the Meng equation is σ_Y, the standard deviation of Y.

When $\rho_{R,Y} \neq 0$, the sampled mean will be non-zero unless $n = N$ (meaning the sample is the entire population) or $\sigma_Y = 0$ (meaning the value of Y is the same for everyone in the population), neither of which are interesting polling contexts.

This is an identity so even when the expected value of $\rho_{R,Y} = 0$ there will be some error (as in the case of random sampling). However as we move to non-random sampling we can expect the realized correlation of R and Y to grow. The larger $\rho_{R,Y}$, the larger the sampling error, the exact magnitude of which will interact with the other terms.

The most explosive implication of the Meng equation emerge from the interaction of the first two terms. When there is MNAR (meaning there will be specific reason to expect $\rho_{R,Y} \neq 0$ because R depends on Y), the actual error depends on the total population. This result is shocking to modern polling sensibilities but is vital to appreciate in the context of non-random sampling.

We can construct a simple two country world to elaborate on how this works. Suppose that our study variable is covid rates and, for the purposes of our example, that covid rates are the same in both countries. One country is huge (China, perhaps) and the other is small (Luxembourg perhaps). If we *randomly* sampled 1,000 people in each country we could produce estimates with the same precision for each country, despite their massive population differences.

What happens if we are dealing with a *non-random* sample of 1,000 people in each country? Suppose for simplicity that people's eagerness for testing is simply a function of their symptoms and that people with more symptoms are more likely to have covid. This creates MNAR sampling because opting into the sample will be associated with higher expected values of our study variable.

In China we will get the 1,000 sickest people. They will be really sick, as they will be in something like the top 0.00001 percentile. In Luxembourg we will also get the 1,000 sickest people, but you don't have to be as sick to get into this set as you would in a much bigger country. This means that the 1,000 sickest people in Luxembourg will be in roughly the top 0.2 percentile; still very sick relative to the population, but not as skewed as in China. In short, MNAR data will produce an error proportional to the population size for a given sample size.

(Note that true random samples are virtually unheard of given non-response among those who are randomly contacted. The actual practice of probability samples can be described as random contact, defined as surveys in which people are randomly contacted even as the response among those contacted may be non-random. Random contact surveys can violate MAR, but nonetheless have strong virtues. Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Bailey (2023) show how survey error in random contact surveys is proportional to the response rate rather than to the population size.)

MAR violations in non-probability sampling lead to errors that are proportional to population size. To use Box's metaphor, this is where the tigers are. Hence as we pursue Wu's exhortation for more coherence in how we evaluate new forms of polling, we should aim to agree on a framework that encompasses the possibility of MAR violations rather than a framework that assumes away this problem.

4. What to do about MAR violations?

Wu follows much of the literature in shying away from MNAR models. Part of the basis for this is a perception that MNAR non-response is essentially intractable. For example, Wu notes somewhat pessimistically that "it is well understood that the MAR assumption cannot be tested using the sample data itself" (page 302) and that "the biased nature of non-probability samples cannot be corrected by using the sample itself" (page 284).

In terms of guidance for survey researchers concerned about violations of MAR, Wu offers only a modest test, which basically consists of finding another variable that is similar to the study variable but that is available for the whole population. If only it were that easy! Generations of pollsters have scoured data for such variables and yet continue to worry about MNAR, especially when response is non-random.

Wu's framing understates what we can do about MAR violations. These efforts will require assumptions, of course, but at least we can relax the severe assumption of MAR. The connection to the earlier points is key: since we are going to need assumptions, it is important that we have a framework for thinking about which ones are the most consequential so that we can focus our efforts appropriately. The Meng equation highlights how MAR violations play a central role in creating error in a non-probabilistic sampling world and therefore we should do whatever we can to address that issue.

A widely known example of a model that can tackle MNAR data is the Heckman (1979) selection model. This model allows for – and even estimates the magnitude of – the MAR violations. It is not without problems, of course. As a practical matter it requires an exclusion restriction (an assumption that one or more variables affect response but not the study variable) and many modern scholars are understandably cautious about the Heckman model's strong parametric assumption.

Scholars have made considerable progress beyond the Heckman model in dealing with MAR violations (Bailey, 2023). The parametric assumption is easy to relax via copula functions (Gomes, Radice, Brenes and Marra, 2019). If we are interested in studying determinants of Y, there is a substantial and growing literature applying highly flexible control functions for MNAR contexts (Das, Newey and Vella, 2003; Liu and Yu, 2022). And if we can identify variables that affect propensity to respond but not the outcome of interest, multiple methods will model and offset MNAR sampling (Peress, 2010; Sun, Liu, Miao, Wirth, Robins and Tchetgen-Tchetgen, 2018).

5. Conclusion

Wu's paper ably and usefully summarizes the state of the literature of analysis of non-probability survey data under the assumption of MAR. He also highlights a critical need for the field to coalesce around a more coherent framework to evaluate these and other polling innovations.

In this note, I build off Wu's work to propose a framework that not only encompasses the MAR models analyzed by Wu, but MNAR models as well, as the violation of the MAR assumption is something particularly relevant and harmful for non-probability surveys.

References

Bailey, M.A. (2023). *Polling at a Crossroads: Rethinking Modern Survey Research*, Cambridge University Press – under contract.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, 600, 695-700.

Das, M., Newey, W.K. and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70, 33-58.

Gomes, M., Radice, R., Brenes, J.C. and Marra, G. (2019). Copula selection models for nongaussian outcomes that are missing not at random. *Statistics in Medicine*, 38, 480-496.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-162.

Liu, R., and Yu, Z. (2022). Sample selection models with monotone control functions. *Journal of Econometrics*, 226, 321-342.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (1): Law of large populations, big data paradox, and the 2016 presidential election. *The Annals of Applied Statistics*, 12, 685-726.

Peress, M. (2010). Correcting for survey nonresponse using variable response propensity. *Journal of the American Statistical Association*, 105, 1418-1430.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J. and Tchetgen-Tchetgen, E.J. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28, 1965-1983.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 319-329
Statistics Canada, Catalogue No. 12-001-X

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

Michael R. Elliott[1]

Abstract

This discussion attempts to add to Wu's review of inference from non-probability samples, as well as to highlighting aspects that are likely avenues for useful additional work. It concludes with a call for an organized stable of high-quality probability surveys that will be focused on providing adjustment information for non-probability surveys.

Key Words: Pseudo-weighting; Propensity score; Doubly-robust estimation; Sensitivity analysis.

1. Introduction

Thanks to Dr. Changbao Wu for an excellent review of the previous work and open issues for statistical inference from non-probability samples. Given the large and rapidly developing work in this area, Dr. Wu was understandably unable to cover all of it; my own understanding has blinders as well but I will touch on a few additional approaches that relate to topics he considered. I will also discuss the issue of modeling versus weighting for different inferential targets, and use his discussion and conclusions to highlight the critical importance of probability samples – in particular high-quality studies that focus on estimation of relevant covariates – to improve inference for the profusion of non-probability samples used as replacements for traditional probability samples in many research and official statistics settings. To avoid notation confusion, all notation will follow that of Wu, except where new notation is required.

Section 2 reviews additional approaches to combining data from probability and non-probability surveys. Section 3 briefly reviews the issue of weighting versus modeling when adjusting non-probability survey data. Section 4 reviews some recent developments in sensitivity analyses of standard assumptions for adjusting non-probability survey data using probability survey data. Section 5 concludes with call to systematically design a set of probability surveys with the explicit purpose of adjusting non-probability surveys.

2. Additional approaches to combining data from probability and non-probability surveys

Dr. Wu's paper follows the general prescription of 1) using model estimation and subsequent calibration to probability-sample-estimated covariate distributions, 2) developing propensity score estimates based on discrepancies between the probability- and non-probability sample data, and 3) doubly-

1. Michael R. Elliott, Department of Biostatistics, University of Michigan, M4124 SPH II, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu.

robust methods that combine 1) and 2) in a manner such that only one of the two underlying models needs to be correct.

2.1 Propensity score estimators

Rivers (2007) appears to have been the first to suggest estimating propensity score using logistic regression with membership in the non-probability sample as the outcome and taking the reciprocal of the resulting propensity scores to use as inclusion weights. This approach was formalized further in Valliant and Dever (2011). Separately, using simple results from Bayes' theorem and discriminant analysis first described in Elliott and Davis (2005), Elliott, Resler, Flannagan and Rupp (2010) and Elliott (2013) developed a somewhat different estimator of the form

$$\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha}) = \hat{P}(i \in S_A) \propto P(i \in S_B) \frac{\hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})}{\hat{P}(i \in S_B \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})}. \quad (2.1)$$

$\hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$ can be obtained using logistic regression, or using one of the suite of machine learning-type approaches such as support vector machines (Soentpiet, 1999), targeted maximum likelihood estimation (Van Der Laan and Rubin, 2006), or Bayesian Additive Regression Trees (BART) (Chipman, George and McCulloch, 2010), and $\hat{P}(i \in S_A \mid i \in S_B \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$ obtained as $1 - \hat{P}(i \in S_A \mid i \in S_A \text{ or } i \in S_B, \mathbf{x}_i, \boldsymbol{\alpha})$. In principle $P(i \in S_B) = 1/d_i^B$ is known since sampling probabilities are known for all elements of the population, including those in the non-probability sample, but in practice analysts with access only to public use data may have to estimate this as well. (In addition, d_i^B may include calibration and non-response adjustments that are not known for the non-probability sample elements.) This last point is critical as use of the probability sample to develop propensity scores using only the discrepancies between the non-probability sample and the probability sample will be biased unless the probability sample used an equal probability (epsem) design, as noted by Wu.

In contrast, Chen, Li and Wu (2020) shows that using a pseudo-likelihood approach to estimating $\hat{\pi}_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ directly from the population likelihood for the indicators $I(i \in S_A)$ as a function of $\mathbf{x}_i$ yields an estimator that does not require $P(i \in S_B)$ for elements in the non-probability sample under the restriction that $\pi_i^A(\mathbf{x}_i, \boldsymbol{\alpha})$ follows a generalized linear model with a canonical link, i.e., logistic regression.

(None of these approaches actually has the correct intercept to obtain a true propensity score; however, as noted in Wu, weighted estimation usually uses Hájek-type estimators [using weights to estimate a population total for denominators; Hájek, 1971] so that propensity scores estimated up to a normalizing constant are sufficient.)

2.2 Doubly-robust estimators

If inference is focused on a particular variable Y available only in the non-probability sample, we can return to the model-assisted estimators that date back to Cassel, Särndal and Wretman (1976), which posit a model for the expectation $E(y_i \mid \mathbf{x}_i) = m_i$. Combining this with propensity score estimates of the

probability of being in the non-probability sample (which we are treating as an "unknown probability sample" – more about this under Assumptions below) yields estimators of the form

$$\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}+\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i \tag{2.2}$$

corresponding to $\hat{\mu}_{\mathrm{DR}2}$ of (4.11) in Wu. The intuition is that any bias due to the model misspecification in estimation of m_i in $\frac{1}{\hat{N}^B}\sum_{i\in S_B}d_i^B\hat{m}_i$ will be equal to and opposite in sign of $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ if the model for π_i^A is correctly specified. Conversely, if the model for π_i^A is misspecified but m_i is correctly specified, $y_i-\hat{m}_i$ will be iid with mean zero and consequently $\frac{1}{\hat{N}^A}\sum_{i\in S_A}\frac{y_i-\hat{m}_i}{\hat{\pi}_i^A}$ will also have mean 0, yielding an unbiased estimator. Chen, Valliant and Elliott (2019) used LASSO for prediction in combination with generalized regression estimators (McConville, Breidt, Lee and Moisen, 2017) when $\mathbf{X}$ is of high dimension. As Wu notes, Wu and Sitter (2001) show the equivalence between GREG applied to predicted values and DR estimators of the form in (2.2), which indicates that the Chen et al. (2019) approach was equivalent to (2.2) with LASSO estimation for m_i and an assumption of simple random sampling for the non-probability sample.

A disadvantage of using (2.1) as opposed to Chen et al. (2020) as the estimator of π_i^A, and thus of d_i^A, is the requirement that the probability sample weights d_i^B be known or at least estimated for the non-probability sample. An advantage of using (2.1), is that non-linear models and machine learning methods can be used in estimation. Rafei, Flannagan and Elliott (2020) uses BART to estimate both m_i and π_i^A, reducing the impact of potential model misspecification. Simulations showed considerable improvement in bias and variance reduction over the method of Chen et al. (2020) when the linear models is misspecified. Variance estimation can proceed by adapting Rubin's multiple imputation rules: from M independent draws from BART, the mean of the variances computed treating the draw of d_i^A as known using standard complex sample design estimators and added to $\frac{M+1}{M}$ times the variance of the point estimates computed across the draws of d_i^A yield an approximately unbiased variance estimator.

An alternative approach to doubly-robust estimation uses the fact that the propensity score is the coarsest possible "balancing score" that contains all of the information about the association between the sampling indicator and the outcome of interest. This has led to the development of mean estimators that use smooth functions of weights to produce consistent estimators that can be more efficient when weights are highly variable or only weakly related to the outcome (Elliott and Little, 2000; Zheng and Little, 2005). Zhou, Elliott and Little (2019) extended this idea into the causal inference setting in non-randomized settings, in which probability of assignment to a treatment or exposure (propensity score) is estimated as a function of covariates $P_Z\left(\mathbf{x}_i,\boldsymbol{\alpha}\right)$ using logistic regression, and then non-observed potential outcomes Y^z under treatment arm $z_i'\neq z_i$ for observed treatment z_i are imputed from

$$Y_i^Z\sim N\left(s\left(\hat{P}_Z^*(\mathbf{x}_i,\hat{\boldsymbol{\alpha}})\,\middle|\,\boldsymbol{\theta}_Z\right)\right)+g_Z\left(\hat{P}^*(\mathbf{x}_i,\hat{\boldsymbol{\alpha}}),\mathbf{x}_i\,\middle|\,\boldsymbol{\beta}_Z\right),\sigma^2 \tag{2.3}$$

where P^* is the logit transformation of P, $s(\hat{P}_Z^* \mid \boldsymbol{\theta}_Z)$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996) of propensity, and $g_Z(\hat{P}^*, \mathbf{x}_i \mid \boldsymbol{\beta}_Z)$ is a general function of covariates including the propensity scores. The resulting estimator is doubly robust in the sense that if either $P_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ or $E(Y^z) = g_Z(\hat{P}^*, \mathbf{x}_i \mid \boldsymbol{\beta}_Z)$ is correctly specified, $Y^{(z)}$ will be approximately unbiased; see Zhang and Little (2009). This can be implemented in the non-probability setting by replacing $\hat{P}_Z(\mathbf{x}_i, \boldsymbol{\alpha})$ in the mean model for (2.3) with $\hat{\pi}_i^A$ estimated using (2.1) to obtain a draw of $Y_i^{(b)}$. (Note this requires obtaining $\hat{\pi}_i^A$ for the probability sample elements requiring prediction.) Inference can proceed by obtaining $b = 1, \ldots, B$ draws from the posterior distribution of the estimated population quantity of interest, e.g., for the population mean

$$Y^{(b)} = \frac{\sum_{i \in S_R} N_i^{(b)} Y_i^{(b)} + \sum_{i \in S_A} \left(y_i - Y_i^{(b)} \right)}{N}$$

where now $N_i^{(b)}$ is a estimate of the population represented by the weight d_i^R obtained from a finite population Bayesian bootstrap (Little and Zheng, 2007); more complete FBPP extensions to complex sample designs that include clustering and stratification are available in Dong, Elliott and Raghunathan (2014).

As in the estimation of (2.1), the non-parametric (spline) component of (2.3) can be replaced with other machine-learning estimators; see Chapter 4 of Rafei (2021) for implementation using Gaussian processes. Also, extensions to non-normal models are direct, although not necessarily computational easy.

2.3 Poststratified estimators

Wu also describes the use of poststratified estimators in the context of quota sampling, which is not only a very old form of non-probability sampling but indeed the standard before Neyman made the case for stratified random sampling (Neyman, 1934). Wu's Section 5 suggests a robust alternative to the propensity score estimates obtained by ordering observations in the probability sample by $\hat{\pi}_i$, stratifying into K strata based on this ordering, and computing the predicted proportion of the population belonging to the k^{th} stratum as proportion of the sample weights W_k in this stratum using the probability sample, with

$$\hat{\mu}_{\text{PST}} = \sum_k \hat{W}_k \bar{y}_k \tag{2.4}$$

where $\bar{y}_k$ is the mean within the k^{th} stratum in the non-probability sample. Wu notes the tradeoff between choosing K to be large enough to retain homogeneity within units but small enough to obtain stable estimates of $\bar{y}_k$, suggesting 30 as the old "rule of thumb" for "large [enough] sample sizes". I would add that a more formal approach discussed in Little (1986) suggests a method to generate strata (there in the context of non-response adjustment) that minimizes mean square error by maximizing the

between-stratum-to-within-stratum variance. It would seem such an approach would be appropriate to consider in the non-probability post-stratified estimator as well.

A more direct approach to obtain estimates using a post-stratified type estimator is multilevel regression and poststratification (Wang, Rothschild, Goel and Gelman, 2015; Downes and Carlin, 2020). Here only data from the non-probability sample is used in the outcome model:

$$E\left(Y_{k[i]}\right) = \beta_0 + \mathbf{x}_k^T\boldsymbol{\beta} + \sum_j a_{l[k]}^j \tag{2.5}$$

where $k = 1,\ldots,K$ indexes the poststratum developed from $j = 1,\ldots,J$ variables, $a_{l[k]}^j \sim N(0,\sigma_j^2)$ for $l = 1,\ldots,L_j$ and $l[k]$ maps the postratum cell k to the appropriate category l of variable j. The poststratifed estimator is still given by (2.4) with $\hat{W}_k$ now replaced with known population totals W_k; posterior inference is obtained though posterior draws of β_0, $\boldsymbol{\beta}$, and $a_{l[k]}^j$ to obtain a draw of

$$\hat{\mu}_{\text{PST}}^{(b)} = \sum_k W_k \left[\frac{1}{n_k}\sum_{i\in k}\left(\beta_0^{(b)} + \mathbf{x}_k^T\boldsymbol{\beta}^{(b)} + \sum_j a_{l[k]}^{j(b)}\right)\right].$$

Though not technically doubly-robust, it has been shown to work well in some applications where J is large enough to capture all of the important discrepancies between the probability and non-probability sample, and the non-probability sample is sufficiently large to allow reasonably accurate estimation of $a_{l[k]}^j$. In the absence of known joint distributions of a high dimensional $\mathbf{X}$, this approach has the weakness of relying on estimated distributions, which are unstable. A possible alternative might be replace the simple $\bar{y}_k$ with (2.5) in Wu's poststratified estimator (2.4), using the fact that the sampling weights d_i^R summarize the information about $\mathbf{X}$ in the probability sample similar to that of the propensity score for non-probability sample.

3. Weighting vs. modeling for the general user

Wu's paper and the above addendums tend to follow the long-trodden path regarding weighting versus modeling in the finite population inference setting, dating back at least to Hansen, Madow and Tepping (1983). In thinking about this choice I believe it is important to distinguish between models used to derive so-called descriptive parameters – in the sense of Kalton (1983) – and models that are of interest in and of themselves, so-called analytic parameters in regression models, latent classes analysis, etc. For the former distinguishing a descriptive target of interest Y from potential modeling covariates $\mathbf{X}$ has the advantage of creating doubly-robust estimators that are targeted to a single descriptive parameter. This also requires assumptions such as A1 in Section 2.1 (propensity score does not depend on Y conditional on $\mathbf{X}$). When models themselves are the targets of interest, it may be that developing weights via propensity scores to account for selection bias and, as Wu notes, employing standard weighted estimating equations may be the most sensible choice, since typically a wide number of models may be considered. This comes at the cost

of double robustness, since there is usually no attempt to model the analytic parameter directly. Developing ways to extend double-robustness into a broader class of model parameter estimates may be a fruitful exercise.

4. Unverifiable assumptions: Recent developments in sensitivity analysis

Wu provides four key assumptions required to correct for selection bias in non-probability surveys using data from probability surveys: they can be roughly summarized as "selection at random" or SAR (covariates in the non-probability sample explain the probability of selection in the non-probability sample); "positivity" (all elements in the population have a non-zero probability of selection into the non-probability sample); "independence" (elements are selected independently into the non-probability sample); and "common covariates" (there exists a probability survey with covariates whose subset matched the covariates required for the MAR assumption to hold). It might be worth noting that the first two assumptions basically require the non-probability survey to be a probability survey "in disguise" – that is, there really are non-zero probabilities of selection into the non-probability survey for all elements in the population, but we as analysts just do not know what they are.

In practice neither of these assumptions probably hold precisely. Some recent work has focused on the failure of the first, the SAR assumption. Some existing measures borrowed from the non-response literature have been repurposed here: for example, the R-indicator measure (Schouten, Cobben and Bethlehem, 2009), which in this context is the measure of the variability in the probabilities of selection in the non-probability sample:

$$\hat{R} = 1 - 2\sqrt{\frac{1}{n_a - 1}\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A - \sum_{j=1}^{n_A}\hat{\pi}_j^A / n_a\right)^2}$$

$\hat{R}$ can range between 0 and 1, where 1 is achieved when probabilities of selection are constant – suggesting something akin to a simple random sample, with less chance for selection bias – and 0 – suggesting all elements are either included with probability 1 or 0, maximizing the risk of selection bias.

Of course, in the absence of the outcome Y in the probability sample, there is no way to directly assess selection bias. Hence recent work has extended Andridge and Little (2011), which develops a sensitivity analysis using a pattern-mixture model, wherein selection into non-probability sample is allowed to depend entirely on a scalar reduction to the covariates $\mathbf{X}$, entirely on the outcome Y, or some convex combination thereof. Little, West, Boonstra and Hu (2020), Andridge, West, Little, Boonstra and Alvarado-Leiton (2019), and West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton (2021) consider sensitivity to this assumption in the estimation of the mean of a normally distributed variable, the mean of a binary outcome, and the regression parameters in a linear regression model, respectively, in non-probability samples. By varying the convex mixing parameter ϕ, sensitivity to the SAR assumption

can be assessed. Boonstra, Little, West, Andridge and Alvarado-Leiton (2021) finds that these "standard measures of bias" (SMB) compare favorably with alternatives such as $\hat{R}$ in a simulation study. An important point to note is that the methods that extend Andridge and Little (2011) do not depend on assumption of common covariates in a probability sample. This suggests that methods that use information available in the probability sample to assess SAR are an open area for development.

The second assumption – positivity – is also unlikely to exist precisely in many practical settings. My own work in this area has focused on naturalistic driving studies, which typically involve convenience samples in a limited geographical area: for example, the Second Strategic Highways Research Program (SHRP2) recruited drivers in six specific geographic regions across the United States (Transportation Research Board (TRB) of the National Academy of Sciences, 2013). This corresponds to the second scenario given by Wu in Section 7.2, where only a subpopulation has any chance of being selected into the non-probability sample, which as he notes has "no simple fix". Following his notation of D providing an indicator of membership in the subpopulation, it would seem that if $D_i \perp \mathbf{X}_i, Y_i \mid \pi_i^A$ – that is, if the distribution of $\mathbf{X}, Y$ is the same for $D=0$ and $D=1$ after weighting for π_i^A within the $D=1$ stratum – then lack of positivity would have no impact on inference. This is likely a tall order in the most general settings but might be reasonably well approximated if the analysis of interest involves a subset of $\mathbf{X}, Y$ that is only weakly associated with D even before adjustment.

Finally, regarding the fourth assumption – existence of a probability sample with available $\mathbf{X}$ – I very much second Wu's observation that methods to take advantage of multiple probability surveys need more development. However, it remains more likely that a researcher will struggle to find a single probability sample with sufficient covariates than struggle with a surfeit of options (Wu's "rich person's problem"). To this end I will conclude with a call to action by the survey community.

5. Probability sampling in the 21st century: Now more than ever

I learned statistics, and particularly survey statistics, near the end of the 20th century, when probability sampling was the unchallenged touchstone of survey design. I was first introduced to the problem of making inference from non-probability samples in the late 00's in the context of injury analysis using Crash Injury Research (CIREN) data, where analysts were treating a highly-restricted sample of individuals in passenger vehicle crashes as if they were a random sample of crash victims and consequently finding non-sensible results (Elliott et al., 2010). About the same time web surveys were exploding in popularity and survey statisticians were somewhat at a loss as to how to make inference from such data. I will admit to a rather paternalistic attitude at the time – I almost avoided trying to do research in this area because I thought it would only encourage "bad behavior" regarding sample design. I did not think I could single-handedly stop it, but I did not want to participate in what I perceived as the downgrading of science. I came to recognize, however, that many of these new data sources have advantages beyond what can be achieved through the traditional probability sample, certainly within

limited budgets. This is above and beyond the increasing challenges to implementing probability surveys, especially in general populations, due to non-response, lack of adequate sampling frames, etc.

However, I remain concerned that the idea that we have developed methods to deal with the limitations of non-probability surveys means that probability sampling is passe is becoming entrenched among scientists and policy makers with limited statistical training, despite efforts like those of Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) and Marek, Tervo-Clemmens, Calabro et al. (2022). However, as Wu's review notes, the absence of probability samples unmoors the non-probability sample from the possibility of even partial calibration or other adjustment approaches (although sensitivity analyses such as those SMB approaches noted above do not require benchmarking probability samples). Hence I believe 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. Some of these obviously already exist – the US Census' American Community Survey and the National Center for Health Statistics National Health Interview Survey premier among them – but going forward I believe it would be valuable for statistical agencies to explicitly coordinate around the need for high quality probability surveys to serve a role as analytic partners to the non-probability survey world rather than just as stand-alone products. This means thinking carefully about important covariates across a variety of public health and social science roles in which survey data play a role. Choices will have to be made given limited budget constraints, and at the same time provisions should be made for sufficient funding to retain the quality needed for adjustment. Finally, while some methods do not require microdata and thus can use summary measures such as those avaiable in the American Communities Survey, other will require such data, which likely means new areas of research to be explored in the fields of privacy and confidentiality research as applied to the combining of data from probability and non-probability surveys.

References

Andridge, R.R., and Little, R.J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27, 153-180.

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. and Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society*, C68, 1465-1483.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600, 695-700.

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, J.K.T., Valliant, R.L. and Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society*, 68, 657-681.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4, 266-298.

Dong, Q., Elliott, M.R. and Raghunathan, T.E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40, 1, 29-46. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014001/article/14003-eng.pdf.

Downes, M., and Carlin, J.B. (2020). Multilevel regression and poststratification as a modeling approach for estimating population quantities in large population health studies: A simulation study. *Biometrical Journal*, 62, 479-491.

Eilers, P.H., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-121.

Elliott, M.R. (2013). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6).

Elliott, M.R., and Davis, W.W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: Combining data from the Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. *Journal of the Royal Statistical Society*, C54, 595-609.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Elliott, M.R., Resler, A., Flannagan, C.A. and Rupp, J.D. (2010). Appropriate analysis of CIREN data: Using NASS-CDS to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis and Prevention*, 42, 530-539.

Hájek, J. (1971). Comment on a paper by D. Basu. *Foundations of Statistical Inference*, 236.

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

Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, 175-188.

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

Little, R.J.A., and Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8, 1-20.

Little, R.J.A., West, B.T., Boonstra, P.S. and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8, 932-964.

Marek, S., Tervo-Clemmens, B., Calabro, F.J. et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, in press.

McConville, K.S., Breidt, F.J., Lee, T. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5, 131-158.

Neyman, J. (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.

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. University of Michigan Thesis. Accessible at https://www.overleaf.com/project/6228db145a47be05f8da3777.

Rafei, A, Flannagan, C.A.C. and Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8, 148-180.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meetings*. Available at https://static.texastribune.org/media/documents/Rivers_matching4.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Soentpiet, R. (1999). *Advances in Kernel Methods: Support Vector Learning*. Boston: MIT Press.

Transportation Research Board of the National Academy of Sciences (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset.*

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods and Research*, 40, 105-137.

Van Der Laan, M.J., and Rubin, D.R. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31, 980-991.

West, B.T., Little, R.J.A., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The Annals of Applied Statistics*, 15, 1556-1581.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zhang, G., and Little, R.J.A. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65, 911-918.

Zheng, H., and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.

Zhou, T., Elliott, M.R., and Little, R.J.A. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114, 1-19.

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.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 339-360
Statistics Canada, Catalogue No. 12-001-X

Comments on "Statistical inference with non-probability survey samples" – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples

Xiao-Li Meng[1]

Abstract

Non-probability samples are deprived of the powerful *design probability* for randomization-based inference. This deprivation, however, encourages us to take advantage of a natural *divine probability* that comes with any finite population. A key metric from this perspective is the *data defect correlation (ddc)*, which is the model-free finite-population correlation between the individual's sample inclusion indicator and the individual's attribute being sampled. A data generating mechanism is equivalent to a probability sampling, in terms of design effect, if and only if its corresponding *ddc* is of $N^{-1/2}$ (stochastic) order, where N is the population size (Meng, 2018). Consequently, existing valid linear estimation methods for non-probability samples can be recast as various strategies to miniaturize the *ddc* down to the $N^{-1/2}$ order. The quasi design-based methods accomplish this task by diminishing the variability among the N inclusion propensities via weighting. The super-population model-based approach achieves the same goal through reducing the variability of the N individual attributes by replacing them with their residuals from a regression model. The doubly robust estimators enjoy their celebrated property because a correlation is zero whenever one of the variables being correlated is constant, regardless of which one. Understanding the commonality of these methods through *ddc* also helps us see clearly the possibility of "double-plus robustness": a valid estimation without relying on the full validity of either the regression model or the estimated inclusion propensity, neither of which is guaranteed because both rely on *device probability*. The insight generated by *ddc* also suggests *counterbalancing sub-sampling*, a strategy aimed at creating a miniature of the population out of a non-probability sample, and with favorable quality-quantity trade-off because mean-squared errors are much more sensitive to *ddc* than to the sample size, especially for large populations.

Key Words: Data defect index; Design probability; Divine probability; Device probability; Design-based inference; Model-assisted survey estimators; Non-response bias.

1. Distinguish among design, divine, and device probabilities

1.1 What can statistics/statisticians say about non-probability samples?

Dealing with non-probability samples is a delicate business, especially for statisticians. Those who believe statistics is all about probabilistic reasoning and inference may question if statistics has anything useful to offer to the non-probabilistic world. Whereas such questioning may reflect the inquirers' ignorance about or even hostility towards statistics, taking the question conceptually, it deserves statisticians' introspection and extrospection. What kind of probabilities are we referring to when the sample is non-probabilistic? The entire probabilistic sampling theory and methods are built upon the randomness introduced by powerful sampling mechanisms, which then yields the beautiful designed-based inferential framework without having to *conceive* anything else is random (Kish, 1965; Wu and Thompson, 2020; Lohr, 2021). When that power – and beauty – is taken away from us, what's left for statisticians?

1. Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, MA 02138. E-mail: meng@stat.harvard.edu.

A philosophical answer by some statisticians would be to dismiss the question altogether by declaring that there is no such thing as probability sample in real life. (I was reminded by Andrew Gelman about this sentiment when I sought his comments on this discussion article. See https://statmodeling.stat.columbia.edu/2014/08/06 for a related discussion.) By the time the data arrive at our desk or disk, even the most carefully designed probability sampling scheme would be compromised by the imperfections in execution, from (uncontrollable) defects in sampling frames to non-responses at various stages and to measurement errors in the responses. In this sense, the notion of probability sample is always a theoretical one, much like efficient market theory in economics, which offers a mathematically elegant framework for idealization and for approximations, but should never be taken literally (e.g., Lo, 2017).

The timely article by Professor Changbao Wu (Wu, 2022) provides a more practical answer, by showcasing how statisticians have dealt with non-probability samples in the long literature of sample surveys and (of course) observational studies, especially for causal inference; see Elliott and Valliant (2017) and Zhang (2019) for two complementary overviews addressing the same challenge. To better understand how probability theory is useful for non-probability samples, it is important to recognize (at least) three types of probabilistic constructs for statistical inference, as listed in Section 1.2. Non-probability samples take away only one of the three, and as a result, they typically force a stronger reliance on the other two.

With these conceptual issues clarified, the rest sections discuss a unified strategy for dealing with non-probability samples. Section 2 reviews a fundamental identity for estimation error, which has led to the construction of data defect correlation (Meng, 2018). Section 3 then discusses how this construct suggests the unified strategy. Section 4 demonstrates the strategy respectively for the qp and ξp settings in Wu (2022). Section 5 then applies the strategy to the two settings simultaneously to reveal an immediate insight into the celebrated double robustness, as reviewed in Wu (2022). Inspired by the same construct, Section 6 explores *counterbalancing sampling* as an alternative strategy to weighting. Section 7 concludes with a general call to treat probability sampling theory as an aspiration instead of the centerpiece of survey and sampling research.

1.2 A trio of probability constructs

The first of the three named constructs below, design probability, is self-explanatory. It is at the heart of sampling theory and reified by practical implementation, however imperfect the implementation might be. The distinction between the next two, divine probability and device probability, may be more nuanced especially at practical levels. But their conceptual differences are no less important than distinguishing between an estimand and an estimator. Fittingly, the data recording or inclusion indicator, a key quantity in modeling non-probability samples, provides a concrete illustration of all three probabilistic constructs; see the leading paragraph of Section 4.

Design Probability. A paramount concept and tool for statistics – and for general science – is randomized replications (Craiu, Gong and Meng, 2022). By designing and executing a probabilistic mechanism to generate randomized replications, we create probabilistic data that can be used directly for making verifiable inferential statements. Besides probabilistic sampling in surveys, randomization in clinical trials, bootstraps for assessing variability, permutation tests for hypothesis testing, and Monte Carlo simulations for computing are all examples of statistical methods that are built on design probability. Non-probability samples, by definition, do not come with such design probability, at least not an identified one. Hence, the phrase non-probability samples should be understood as a short hand for "samples without an identified design probability construct".

It is worth to remind ourselves, however, that there is a potential for design probabilities to come back in a substantial way especially for large non-probability data sets, such as administrative data, due to the adoption of differential privacy (Dwork, 2008), for example by US Census Bureau (see the editorial by Gong, Groshen and Vadhan, 2022, and the special issue in *Harvard Data Science Review* it introduces). Differential privacy methods inject well-designed random noise into data for the purpose of protecting data privacy while not unduly sacrificing data utility. Like the design probability used for probabilistic sampling, the fact that the noise-injecting mechanism is designed by the data curator, and is made publicly known, renders the transparency that is critical for valid statistical inference by the data user (Gong, 2022). The area of how to properly analyze non-probability data with differential privacy protection is wide open. Even more so is the fascinating area of how to take into account the existing defects in non-probability data when designing probabilistic protection mechanisms for data privacy to avoid adding unnecessary noise. Readers who are interested in forming a big picture of the statistical issues involved in data privacy should consult the excellent overview article by Slavkovic and Seeman (2022) on the general area of "statistical data privacy".

Divine Probability. In the absence of design probability for randomization-based inference, in order to conduct a (conventional) statistical inference, we typically conceptualize that the data at hand is a realization of a generative probabilistic mechanism given by nature or God. (I learned about the term "God's model" during my PhD training, which I took as an expression for faith or something beyond human control, rather than reflecting one's religious belief. The phrase "divine" is adopted here with a similar connotation.) We do so regardless of whether we believe or not that the world is intrinsically deterministic or stochastic (e.g., see David Peat, 2002; Li and Meng, 2021). We need to assume this divine probability primarily because of the restrictive nature of the probabilistic framework to which we are so accustomed. For example, in order to invoke the assumption of missing at random, we need to conjure a probabilistic mechanism under which the concept "missing at random" (Rubin, 1976) can be formalized. As Elliott and Valliant (2017) emphasized, the quasi-randomization approach, which corresponds to the qp framework of Wu (2022), "assumes that the nonprobability sample actually does have a probability sampling mechanism, albeit one with probabilities that have to be estimated under identifying

assumptions". That is, we replace the design probability by a divine probability that we have faith for its existence, which then typically is treated as the "truth" or at least as an estimand.

Conceptually, therefore, we need to recognize that the assumption of any particular kind of divine probability is not innocent, as otherwise we will not need to rely on our faith to proceed. Nor is it always necessary. Any finite population provides a natural histogram for any quantifiable attributes or a contingency table for any categorizable attributes of its constituents, and hence it induces a divine probability without referencing any kind of randomness, conceptualized or realized, *if our inferential target is the finite population itself* (not a super-population that generates it, for example). The empirical likelihood approach takes advantage of this natural probability framework, which also turns out to be fundamental for quantifying data quality via data defect correlation (see Meng, 2018). The same emphasis was made by Zhang (2019), whose unified criterion was based on the same identity for building data defect correlation; see Section 2 below.

Device Probability. By far, most probabilities used in statistical modeling are devices for expressing our belief, prior knowledge, assumptions, idealizations, compromises, or even desperation (e.g., imposing a prior distribution to ensure identifiability since nothing else works). Whereas modeling reality has always been a key emphasis in the statistical literature, we inevitably must make a variety of simplifications, approximations, and some times deliberate distortions in order to deal with practical constraints (e.g., the use of variational inference for computational efficiency; see Blei, Kucukelbir and McAuliffe (2017)). Consequently, many of these device probabilities do not come with a requirement of being realizable, or even coherent mathematically (e.g., the employment of incompatible conditional probability distributions for multiple chain imputation; see Van Buuren and Oudshoorn (1999)). Nor are they easy or even possible to be validated, as Zhang (2019) investigated and argued in the context of non-probability sampling, especially with the superpopulation modeling approach, which corresponds to the ξp framework of Wu (2022). Nevertheless, device probabilities are the workhorse for statistical inferences. Both quasi-randomization approach and super-population modeling rely on such device probabilities to operate, as shown in Wu (2022) and further discussed in Sections 4-5 below. The lack of design probability can only encourage more device probabilities to make headway. To paraphrase Box's famous quote "all models are wrong, but some are useful", all device probabilities are problematic, but some are problem-solving.

1.3 Let's reduce "Garbage in, package out"

In a nutshell, probabilistic constructs are more needed for non-probability samples than probability ones precisely because of the deprivation of the design probability. Therefore, dealing with non-probability samples is not a new challenge for statisticians. If anything is new, it is the availability of massive amounts of large and non-probabilistic data sets, such as administrative data and social media data, and the accelerated need to combine multiple sources of data, most of which inherently are non-probabilistic because they are not collected for statistical inference purposes (e.g., Lohr and Rao, 2006; Meng, 2014; Buelens, Burger and van den Brakel, 2018; Beaumont and Rao, 2021). Contrary to common

belief, the large sizes of "big data" can make our inference much worse, because of the "big data paradox" (Meng, 2018; Msaouel, 2022) when we fail to take into account the data quality in assessing the errors and uncertainties in our analyses; see Section 6.1.

It is therefore becoming more pressing than ever to greatly increase the general awareness of, and literacy about, the critical importance of data quality, and how we can use statistical methods and theories to help to reduce the data defect. The central concern here goes beyond the common warning about "garbage in, garbage out" – if something is recognized as garbage, it would likely be treated as such (likely, but not always, because as Andrew Gelman reminded me that "many researchers have a strong belief in *procedure* rather than *measurement*, and for these people the most important thing is to follow the rules, not to look at where their data came from"). The goal is to prevent "garbage in, package out" (Meng, 2021), where low quality data are auto-processed by generic procedures to create a cosmetically attractive "AI" package and sold to uninformed consumers or worse, to those who seek "data evidence" to mislead or disinform. Properly handling non-probability samples obviously does not resolve all the data quality issues, but it goes a very long way in addressing an increasingly common and detrimental problem of lack of data quality control in data science.

I therefore thank Professor Changbao Wu for a well timed and designed in-depth tour of "the must-sees" of the large sausage-making factory for processing non-probability samples. It adds considerably more detailed and nuanced exhibitions to the general tour by Elliott and Valliant (2017), which includes excellent illustrations on many forms and shapes of non-probability samples as well as their harms. It also showcases theoretical and methodological milestones for us to better appreciate the millstones displayed in the intellectual tour by Zhang (2019), which challenges statisticians and data scientists in general to understand better the quality, or rather the lack thereof, of the products we produce and promote. Together, this trio of overview articles form an informative tour for anyone who wants to join the force to address the ever-increasing challenges of non-probability data. Perhaps the best tour sequence starts with Elliott and Valliant (2017) to form a general picture, with Wu (2022)'s as the main exhibition of methodologies, and ends with Zhang (2019) to generate deep reflections on some specific challenges. For additional common methods for dealing with non-probability samples, such as multilevel modeling and poststratification, readers are referred to Gelman (2007), Wang, Rothschild, Goel and Gelman (2015) and Liu, Gelman and Chen (2021).

As a researcher and educator, I have been beating similar drums but often frustrated by the lack of time or energy to engage deeply. I am therefore particularly grateful to Editor Jean-François Beaumont for inviting me to help to ensure Professor Wu's messages are loud and clear: data cannot be processed as if they were representative unless the observed data are genuinely probability samples (which is extremely rare); many remedies have been proposed and tried, but many more need to be developed and evaluated. Among them, the concept of data defect correlation is a promising general metric to be explored and expanded, as demonstrated below.

2. A finite-population deterministic identity for actual error

To demonstrate the fruitfulness of the finite-population framework, consider the estimation of the population mean, denoted by $\bar{G}$, of $\{G_i = G(X_i): i \in \mathcal{N}\}$, where $\mathcal{N} = \{1, \ldots, N\}$ indexes a finite population, and the X_i's are data collected on individual i. For each i, let $R_i = 1$ if G_i (or rather X_i) is recorded in our sample, and $R_i = 0$ otherwise; hence the sample size is $n_R = \sum_{i=1}^{N} R_i$. We stress that this is an all-encompassing indicator, which can (and should) be decomposed into $R_i = r_i^{(1)}, \ldots, r_i^{(J)}$, when the data collection consists of J stages (e.g., $r_i^{(1)}$ indicates whether or not the i^{th} individual is sampled, and $r_i^{(2)}$ whether the individual responded or not once sampled).

Let $\{W_i, i \in S\}$ be a set of weights to be determined, where the index set $S = \{i: R_i = 1\}$, such that $\sum_{i \in S} W_i > 0$. Let $\bar{G}_W$ be the weighted sample average, expressible in three ways:

$$\bar{G}_W = \frac{\sum_{i \in S} W_i G_i}{\sum_{i \in S} W_i} = \frac{\sum_{i=1}^{N} R_i W_i G_i}{\sum_{i=1}^{N} R_i W_i} = \frac{\mathrm{E}_I(\tilde{R}_I G_I)}{\mathrm{E}_I(\tilde{R}_I)}, \tag{2.1}$$

where $\tilde{R}_I = R_I W_I$, and E_I is taken with respect to the uniform distribution over the index set $\mathcal{N}$. The first expression in (2.1) simply defines a weighted sample average. With the help of R_i, the second expression turns the sample averages into finite-population averages. This trivial re-expression is fundamental because it explicates the role of R_i in influencing the behavior of $\bar{G}_W$ as an estimator of $\bar{G}$. The third expression reveals a divine probability through I, the finite-population index (FPI) variable, by utilizing the fact that averaging is the same as taking expectation over a uniformly distributed random index I. All finite-population moments then can be expressed via E_I.

In particular, we can express the actual error of $\bar{G}_W$ via the following identity, where the first expression can be traced back to Hartley and Ross (1954), who used it to express biases in ratio estimators. The second expression was given in Meng (2018) with a slightly different (but equivalent) expression:

$$\bar{G}_W - \bar{G} = \frac{\mathrm{Cov}_I(\tilde{R}_I, G_I)}{\mathrm{E}_I[\tilde{R}_I]} = \rho_{\tilde{R},G} \times \sqrt{\frac{N - n_W}{n_W}} \times \sigma_G. \tag{2.2}$$

Here $\rho_{\tilde{R},G} = \mathrm{Corr}_I(\tilde{R}_I, G_I)$ is the *finite-population correlation* between $\tilde{R}_I$ and G_I, σ_G^2 is the finite-population variance of G_I, and n_W is the effective sample size due to using weights (Kish, 1965)

$$n_W = \frac{n_R}{1 + \mathrm{CV}_W^2}, \tag{2.3}$$

with CV_W being the coefficient of variation (i.e., standard deviation/mean) of $\{W_i, i \in S\}$.

The expression (2.2) is an algebraic identity because it holds for any instances of $\{(G_i, R_i W_i), i \in \mathcal{N}\}$. Hence no model assumptions are imposed, not even the assumption that R (or any quantity) is random, echoing the comment by Mary Thompson, as quoted in Wu (2022), that "the sample

inclusion indicator R is a random variable is itself an assumption". The only requirement is that the recorded G_i is unchanged from the G_i's in the target population. (But note this requirement has two components: (1) there is no over-coverage, that is, everyone in the sample belongs to the target population, e.g., no non-eligible voters are surveyed when the target population is eligible voters, and (2) there is no measurement error; extensions to the cases with measurement errors are available, but not pursued in this article.) When we use equal weights, the three factors on the right-hand side of (2.2) reflect respectively (from left to right) data defect, data sparsity, and problem difficulty, as detailed in Meng (2018) and further illustrated in Bradley, Kuriwaki, Isakov, Sejdinovic, Meng and Flaxman (2021) in the context of COVID-19 vaccination surveys.

In particular, when all weights are equal, $\rho_{\tilde{R},G}$ is termed as *data defect correlation* (*ddc*) in Meng (2018) because it measures the lack of representativeness of the sample via capturing the dependence of inclusion/recording indicator on the attributes – the higher the dependence, the more biased the sample average becomes for estimating population averages. With the basic strategies of probabilistic sampling or inverse probability weighting, *ddc* will be zero on average because $\mathrm{E}(W_i R_i) = 1,$ and it is of $O_p(N^{-1/2})$ order because it is essentially an average of N independent terms (Meng, 2018). Our general goal here therefore is to bring down *ddc* to $O_p(N^{-1/2})$ for non-probability samples, which we shall refer to as "miniaturizing *ddc*" because $N^{-1/2}$ is typically a minuscule number in practice.

When we use weights, the first term $\rho_{\tilde{R},G}$ captures the data defect that still exists after the weighting adjustment, since no weights are perfect in practice. Identity (2.2) shows the impact of the weights on both data quality and data quantity. The impact on the *nominal* effective sample size n_W is never positive because $n_W \le n_R$ as seen in (2.3). Incidentally, the exactness of (2.3) reveals that this well-known expression is in fact not an approximation (which is often attributed to Kish (1965)), but an exact formula for the reduction of the sample size due to weighting *if the weighting had no impact on ddc*. However, weighting can have a major positive impact on reducing the overall error by judiciously choosing weights to significantly decrease *ddc*, though apparently at the price of $n_W < n_R$. Of course, this is exactly the aim of the quasi-randomization framework, as discussed below. Most importantly, however, (2.2) leads to a unified insight about the variety of methods reviewed in Wu (2022), including an intuitive explanation of the doubly robust property, which has been receiving increased attention for integrating data sources including both probability and non-probability samples (e.g., Yang, Kim and Song, 2020).

Indeed, Zhang (2019, Section 3.1) used the first expression in (2.2) to define a unified non-parametric asymptotic (NPA) non-informativeness assumption, which requires that the numerator $\mathrm{Cov}_I(\tilde{R}_I, G_I)$ goes to zero, while keeping the denominator $\mathrm{E}_I[\tilde{R}_I]$ positive, as $N \to \infty$. This unification permits Zhang (2019) to evaluate the quasi-randomization approach and regression modeling via a common criterion. The *ddc* framework echoes this unification, as discussed in Section 3 below, with Section 4 stressing the same broad message as emphasized by Zhang (2019). Section 5 harvests another low-hanging fruit of the *ddc* formulation, since it provides an immediate explanation of the celebrated double robustness. Section 6 then ventures into a much harder area of engineering a more representative

sub-sample out of a large non-representative sample, a worthwhile trade-off because data quality is far more important than data quantity (Meng, 2018), as briefly reviewed below.

3. A unifying strategy based on data defect correlation

In the setup of Wu (2022), for each individual i, we have a set of attributes $A_i = \{y_i, \mathbf{x}_i\}$, where y is the attribute of interest, and $\mathbf{x}$ is auxiliary, which is useful in two ways. First, reducing the sampling bias due to non-probability sampling becomes possible when the non-probability mechanism can be (fully) explained by $\mathbf{x}$. Second, by taking advantage of the relationships between y_i and $\mathbf{x}_i$, we can improve the efficiency of our estimation. As a starting point, Wu (2022) assumes that we have two data sources available, which we denote via two recording indicators, R and R^*. The main source of the data is a non-probability sample, where we observe both y_i and $\mathbf{x}_i$ for $i \in S \equiv \{i : R_i = 1\}$, but the recording indicator R_i is determined by a mechanism uncontrolled by any (known) design probability. A second source is (assumed to be) a probability sample, where we observe $\mathbf{x}_i$ only, for $i \in S^* \equiv \{i : R_i^* = 1\}$. This second sample provides information to estimate population auxiliary information that is useful for estimating population quantities about y, such as its mean. Hence this setup is closely related to the setup where $S \cup S^* = \mathcal{N}$; see Tan (2013).

Now for any function $m(\mathbf{x})$, let $z_i = y - m(\mathbf{x}_i), i \in \mathcal{N}$. Clearly we can estimate the population mean $\bar{y}_N = \mathrm{E}_I(y_I)$ via estimating $\bar{z} = \mathrm{E}_I(z_I)$ and $\bar{m} = \mathrm{E}_I[m(\mathbf{x}_I)]$. From the second sample, $\bar{m}$ can be estimated unbiasedly since it involves $\mathbf{x}$ only. We therefore can focus on estimating $\bar{z}$, while recognizing that a more principled approach is to set up a likelihood or Bayesian model to estimate all unknown quantities jointly (Pfeffermann, 2017). Applying identity (2.2) with $G = z$ then tells us that our central task is to choose the weight $\{W_i, i \in S\}$ and/or the m function to miniaturize the *ddc* $\rho_{\tilde{R},z}$. For our current discussion, it is easier to explain everything via the covariance

$$c_{\tilde{R},z} \equiv \mathrm{Cov}_I(\tilde{R}_I, z_I) = \mathrm{Cov}_I(W_I R_I, y_I - m(\mathbf{x}_I)) = \frac{1}{N}\sum_{i=1}^{N} W_i R_i (z_i - \bar{z}) \tag{3.1}$$

instead of the correlation $\rho_{\tilde{R},z}$ because $\mathrm{Cov}_I(\tilde{R}_I, z_I)$ is a bi-linear function in R_I and z_I. However, $\rho_{\tilde{R},z}$, being standardized, is more appealing theoretically and for modelling purposes; see Sections 6 and 7.

The expression in (3.1) tells us immediately how to make it zero in expectations operationally, and in what sense conceptually. For whatever probability we impose on R_i (to be specified in late sections), let $\pi_i = \Pr(R_i = 1 \mid \mathbf{A})$, which we assume will depend on A_i only. Then the linearity of the covariance operator implies that the average covariance with respect to the randomness in R_i is given by

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{A}] = \mathrm{Cov}_I\left(W_I \pi_I, y_I - m(\mathbf{x}_I)\right), \tag{3.2}$$

where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Similarly, if one is willing to posit a joint model for $\{(R_i, y_i), i \in \mathcal{N}\}$ conditioning on $\mathbf{X}$ in the independence form $\Pi_{i=1}^{N} P\left(R_i, y_i \mid \mathbf{x}_i\right)$, then

$$\mathrm{E}[c_{\tilde{R},z} \mid \mathbf{X}] = \mathrm{Cov}_I\left(W_I \pi_I, \mathrm{E}(y_I \mid \mathbf{x}_I) - m(\mathbf{x}_I)\right). \tag{3.3}$$

Very intuitively, one can ensure a zero covariance or correlation between two variables by making either of them a constant. The two choices then would lead to respectively the quasi-randomization approach by making $W_I \pi_I \propto 1$ and the super-population approach by making $\mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)$ a constant (e.g., zero). The fact that either one is sufficient to render zero covariance (under the joint model) yields the double robustness, because it does not matter which one. But clearly these are not the only methods to achieve a zero correlation/covariance or double robustness, an emphasis of Kang and Schafer (2007) in their attempt to demystify the doubly robust approach (Robins, Rotnitzky and Zhao, 1994; Robins, 2000; Scharfstein, Rotnitzky and Robins, 1999). See also Tan (2007, 2010) for discussions and comparisons of an array of estimators, including those corresponding to only the quasi-randomization approach or only the super-population approach, some of them are doubly robust.

Indeed, because formula (2.2) is an identity for the actual error, any asymptotically unbiased (linear) estimators of the population mean must imply its corresponding *ddc* is asymptotically unbiased for zero, and vice versa, with respect to the randomness in R or in $\{R, y\}$. However, it is possible for *ddc* to be asymptotically unbiased for zero, without assuming any model is correctly specified – see Section 5 for an example. (This "double-plus robustness" is different from the "multiple robustness" of Han and Wang (2013), which still needs to assume the validity of at least one of the posited multiple models.) These two observations suggest that any general sufficient and necessary strategy for ensuring asymptotically consistent/unbiased (linear) estimators for the population mean would be equivalent to miniaturizing *ddc*.

As an example of a unified insight that otherwise might not be as intuitive, expression (3.2) suggests that we should include our estimate of π_I as a part of the predictor in the regression model $m(\mathbf{x}_I)$, since that can help to reduce the correlation between $W_I \pi_I$ and $z_I = y_I - m(\mathbf{x}_I)$, especially when we use constant weights W_I. Using $\hat{\pi}_I$ as a predictor for y is generally hard to motivate purely from the regression perspective, especially when we assume y and R are independent given $\mathbf{x}$ (typically a necessary condition to proceed, as discussed in the next section). However, expression (3.2) tells us that for the purpose of estimating the mean of y, it is not absolutely necessary to fit the correct regression model $m(\mathbf{x})$. Rather, it is sufficient to ensure the "residual" z_I is as uncorrelated with $W_I \pi_I$ as I varies. However, it is critically important to recognize that it is not sufficient to ensure zero or small correlation only among the observed data, because $\mathrm{Cov}_I(W_I \pi_I, z_I \mid R_I = 1)$ tells us little about $\mathrm{Cov}_I(W_I \pi_I, z_I \mid R_I = 0)$. In the setting of Wu (2022), our ability to extrapolate from $R_I = 1$ to $R_I = 0$ depends on the availability of the (independent) auxiliary data indexed by $R_I^* = 1$, which allow us to observe some x_I's for which $R_I = 0$.

The strategy of including propensity estimates as a predictor has been found beneficial in related literature. For example, Little and An (2004) included the logit of $\hat{\pi}$ in their imputation model, and

reported the inclusion enhanced the robustness of the imputed mean to the misspecification of the imputation model. The method was further developed and enhanced by Zhang and Little (2009) and by Tan, Flannagan and Elliott (2019), who used the term "Robust-squared" to emphasize the enhanced robustness. In a more recent article on such a strategy for non-probability samples, Liu et al. (2021) emphasized the importance of including the estimated propensity $\hat{\pi}_i$ "as a predictor" in $m(x, \hat{\pi})$ (using notation in this article). Furthermore, in the literature of targeted maximum likelihood estimation (TMLE) for semi-parametric models for dealing with non-probability data (van der Laan and Rubin, 2006; Luque-Fernandez, Schomaker, Rachet and Schnitzer, 2018) (also see Scharfstein et al. (1999); Tan (2010)), the variables $R_I / \hat{\pi}_I$ and $(1 - R_I) / (1 - \hat{\pi}_I)$ are called *clever covariates* and are used in the regression models for y_I. The implementations and theories of TMLE, and the related Collaborative TMLE (van der Laan and Gruber, 2009, 2010), are mathematically more involved than those under finite-population settings as discussed below, but the insights gained from (3.2)-(3.3) can provide us with helpful intuitions on understanding the essence of such methods.

4. Quasi-randomization *or* super-population implementations

In a nutshell, the quasi-randomization approach focuses on making $W_I \pi_I$ a constant variable (induced by FPI I). When our sample is genuinely selected by a probabilistic scheme by design, then $\pi_i = \Pr(R_i = 1 \,|\, \mathbf{x}_i)$, for $i \in \mathcal{N}$, is a design probability, free of y_i, but it can depend on $\mathbf{x}_i$ for example when $\mathbf{x}_i$ includes a stratifying variable. When the design probability is unavailable, we first need to invoke a divine probability. This could be a natural one given by the finite population, such as the propensity $\pi_i = \Pr_I(R_I = 1 \,|\, A_I = A_i)$ induced by FPI, where $A_i = \{y_i, \mathbf{x}_i\}$, or an imagined super-population one such as the R_i's being generated independently from $\text{Ber}(\pi_i)$, where $\pi_i = \Pr(R_i = 1 \,|\, A_i) > 0$. This positivity assumption is necessary if the finite population is pre-specified, or its imposition defines the finite population that can be studied. (This is a practically rather relevant consideration, such as in election polling, where the finite population may not be always pre-specified even theoretically.) Since these divine probabilities are unknown and serve as our estimand, we need to assume some device probabilities, such as via a generalized linear model $\pi_i = g(y_i, \mathbf{x}_i)$ to proceed, even though we don't really believe in any particular choice of g.

For our current discussion, suppose our divine probability is given by the super-population Bernoulli model. Let $n_R = \sum_{i=1}^{N} R_i$, and $\tilde{p}(\mathbf{A}) = \Pr(n_R > 0 \,|\, \mathbf{A}) = 1 - \Pi_{i \in N}(1 - \pi_i)$, where $\mathbf{A} = \{A_i, i \in \mathcal{N}\}$. Because the R_i here is controlled by a divine probability, the sample size n_R is no longer a design variable to be conditioned upon in our replication scheme; it is generally no longer an ancillary statistic. Nevertheless, we should condition on $n_R > 0$, a universal requirement for constructing data-driven estimates for $\bar{G}$. Fortunately this conditioning does not create mathematical complications to the simplicity granted by the independence among $\pi_i, i \in \mathcal{N}$ as functions of A_i. This is because $\tilde{\pi}_i(\mathbf{A}) \equiv \Pr(R_i = 1 \,|\, \mathbf{A}, n_R > 0) = \pi_i / \tilde{p}(\mathbf{A})$, but the normalizing constant $\tilde{p}(\mathbf{A})$ – which depends on

the entire $\mathbf{A}$ – is not relevant for the developments in this article, such as assigning weights that are proportional to $\tilde{\pi}_i^{-1}(\mathbf{A})$.

Consequently, under this divine probability, which corresponds to (the true model for) the q-model setting in Wu (2022), we have for any chosen W_I, by (3.1)

$$\begin{aligned} \mathrm{E}(c_{\tilde{R},z} \mid \mathbf{A}, n_R > 0) &= \mathrm{Cov}_I(W_I \mathrm{E}[R_I \mid \mathbf{A}, n_R > 0], y_I - m(\mathbf{x}_I)) \\ &= \tilde{p}^{-1}(\mathbf{A})\, \mathrm{Cov}_I(W_I \pi_I, y_I - m(\mathbf{x}_I)), \end{aligned} \tag{4.1}$$

where E is with respect to the (unknown) divine probability over R_I (for fixed I). It follows then that, regardless of whether we want to ensure zero expectation in (3.2) or in (4.1), we will impose $W_I \pi_I \propto 1$, that is, $W_I \propto \pi_I^{-1}$, the well-known inverse probability weighting. Therefore, if our postulated model q permits us to reliably capture π_i in reality, then $c_{\tilde{R},z} = O_p(N^{-1/2})$ because it has mean zero (with respect to the divine probability), and it is a weighted average of N essentially independent Bernoulli variables, as seen in (3.1).

This is a randomization oriented approach because it treats the entire finite population attribute values $\mathbf{A}$ as fixed, and the hypothetical replications are generated only by repeated realizations of the recording indicator R_I. Of course, in general, the values of $\{\pi_i, i \in \mathcal{N}\}$ are unknown, and worse they are inestimable from a non-probability sample without further assumptions. To proceed, we pose assumptions such as missing at random, i.e., $\Pr(R_i = 1 \mid A_i) = \Pr(R_i = 1 \mid \mathbf{x}_i)$, and the requirement of an auxiliary sample so that we have some values of $\mathbf{x}_i$ with $R_i = 0$. We also have choices on how to estimate the inclusion propensity $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$, parametrically or non-parametrically. These assumptions, requirements, and estimation methods are all essential for practical implementation, as carefully reviewed and discussed by Wu (2022); also see Tan (2010) for a detailed comparison of various estimation strategies. Nevertheless, the overarching idea of quasi-randomization methods is to choose W_I to free $\tilde{R}_I = W_I R_I$ from I in expectation over the posited hypothetical replications, to regain the freedom guaranteed by probability sampling.

Complementarily, the super-population approaches aim to miniaturize $c_{\tilde{R},z}$ via making the other variable in $c_{\tilde{R},z}$, that is, z_I free of I in expectation, but over a different hypothetical replication scheme. Here the idea is to choose an $m(\mathbf{x}_i)$ that is a good approximation to y_i such that the residual $z_i = y_i - m(\mathbf{x}_i)$ will be zero in expectation conditioning on $\mathbf{x}$. Typically, this is done by considering a joint model for $\{R_i, y_i\}$ given $\mathbf{x}_i$, and with a specific regression model $\xi(y \mid \mathbf{x})$, using the notation in Wu (2022). It is important to recognize that, although we only specify the regression model y_i given $\mathbf{x}_i$, we must include R_i in the replications in order to capture the possible dependence of R_i on the entire $A_i = \{y_i, \mathbf{x}_i\}$, which is the key concern for non-probability samples. Indeed, it is this joint specification that permits the adoption of the missing at random assumption to reduce $P(y_i \mid \mathbf{x}_i, R_i) = P(y_i \mid \mathbf{x}_i)$, which in turn permits us to focus on specifying a single regression model $\xi(y_i \mid \mathbf{x}_i)$ for both observed and unobserved individuals. Therefore, when we write E_ξ, we mean the expectation with respect to

$$P(R_i, y_i \mid \mathbf{x}_i) \;=\; P(R_i \mid \mathbf{x}_i)\, P(y_i \mid R_i, \mathbf{x}_i) \;=\; \pi_i^{R_i}\,(1-\pi_i)^{1-R_i}\,\xi(y_i \mid \mathbf{x}_i), \tag{4.2}$$

where $\pi_i = \Pr(R_i = 1 \mid \mathbf{x}_i)$ is left unspecified, unlike with the quasi-randomization approach.

It follows then that, conditioning on $\mathbf{X} = \{\mathbf{x}_i, i \in \mathcal{N}\}$ and $n_R > 0$, which does not alter $P(y \mid \mathbf{X})$ because y and R are independent given $\mathbf{X}$, we have

$$\mathrm{E}(c_{\tilde{R},z} \mid \mathbf{X}, n_R > 0) \;=\; [\tilde{p}(\mathbf{X})]^{-1}\,\mathrm{Cov}_I\,(W_I \pi_I, \mathrm{E}[y_I \mid \mathbf{x}_I] - m(\mathbf{x}_I)). \tag{4.3}$$

Clearly, (4.3) becomes zero when we choose $m(\mathbf{x}_I) = \mathrm{E}_\xi[y_I \mid \mathbf{x}_I]$ and that the ξ model is (first-order) correctly specified, that is, $\mathrm{E}_\xi[y_I \mid \mathbf{x}_I] = \mathrm{E}[y_I \mid \mathbf{x}_I]$. This summarizes the super-population approach, and it renders $c_{\tilde{R},z} = O_p(N^{-1/2})$ for similar reasons as given for the quasi-randomization framework.

5. Quasi-randomization *and* super-population implementations

Once a joint model for $\{R_i, y_i\}$ is set up, of course we can use it for estimating both π_i and the regression function $m(\mathbf{x})$, each of which is made possible by the availability of the auxiliary probability sample, and the assumption of missing at random. But as shown before, correctly specifying and estimating one of them is sufficient for miniaturizing $c_{\tilde{R},z}$. However, from (4.3), in order for the covariance/correlation to be zero, neither multiplicative correction to π_I via W_I nor the additive adjustment for $\mathrm{E}(y_I \mid \mathbf{x}_I)$ via $m(\mathbf{x}_I)$ need to be correct. All we need is that, after the correction or adjustment, what is left would be uncorrelated with each other. The aforementioned framework of Collaborative TMLE was built essentially on this insight (e.g., see Section 3.1 of van der Laan and Gruber, 2009), though the heavy mathematical treatments in its literature might have discouraged readers to seek such intuitive understanding.

To provide a simple illustration, consider a finite population that is an i.i.d. sample from a super-population model:

$$\mathrm{E}[y \mid x] \;=\; \sum_{k=0}^{3} \beta_k x^k, \quad x \;\sim\; N(0,1). \tag{5.1}$$

The non-probability sample is generated by a mechanism R such that $\Pr(R = 1 \mid y, x) = \pi(|x|)$, that is, it is determined by the magnitude of x only. Suppose we mis-specify the function form for π (e.g., the divine model may not be monotone in $|x|$, but the device model such as the conventional logistic link is), as well the regression model by choosing $m(x) = b_0 + b_1 x + b_2 x^2$. Since x^2 is uncorrelated with x or x^3 under $x \sim N(0,1)$, we know that our least-square estimator for b_2 would still be valid for β_2 even under the mis-specified regression model. This turns out to be sufficient to ensure the asymptotic unbiasedness (as $N \to \infty$) of the following "doubly robust" estimator for $\mu = \bar{y}_N$, the finite-population mean,

$$\hat{\mu}_{+} = \frac{\sum_{i=1}^{N} R_i w(|x_i|)(y_i - \hat{m}(x_i))}{\sum_{i=1}^{N} R_i w(|x_i|)} + \frac{\sum_{i=1}^{N} R_i^* \hat{m}(x_i)}{\sum_{i=1}^{N} R_i^*}, \tag{5.2}$$

where R^* indicates the auxiliary sample (of $\mathbf{x}$ only). Or equivalently,

$$\hat{\mu}_{+} - \bar{y}_N = \frac{\mathrm{Cov}_I\left(R_I w(|x_I|), y_I - \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I w(|x_I|)\right)} + \frac{\mathrm{Cov}_I\left(R_I^*, \hat{m}(x_I)\right)}{\mathrm{E}_I\left(R_I^*\right)}, \tag{5.3}$$

which makes it clearer that any bias in $\hat{\mu}_{+}$ is controlled by the covariance (or correlation) involving R, since the covariance involving R^* is already miniaturized by the assumption that the auxiliary sample is probabilistic (which, for simplicity, is assumed to be a simple random sample).

Here $w(x)$ is any weight function such that $\mathrm{E}_\phi\left[|x|^3 w(|x|)\right] < \infty$, where the expectation is with respect to $x \sim N(0,1)$, and $\hat{m}(x) = b_0 + b_1 x + \hat{\beta}_2 x^2$, with $\hat{\beta}_2$ being the least-square estimator for β_2 from the biased sample, and b_0 and b_1 can be chosen arbitrarily. Because the finite-population covariance/correlation between $\pi(|x_I|) w(|x_I|)$ and x_I^k is $O_p(N^{-1/2})$, for $k = 1$ and $k = 3$, the misfitted parts for π or m do not contribute to the *ddc* (asymptotically) since they are uncorrelated with each other under the super-population model, leading to further robustness going beyond "double robustness". This of course does not mean that we can misfit a model arbitrarily and still obtain valid estimators, but it does imply that having at least one model being correct is a sufficient, but not necessary, condition for the validity of the doubly robust estimators.

It is also worth stressing that, in formatting the regression model, we do not necessarily need to invoke a device probability, e.g., a super-population regression model, because the FPI variable provides a finite-population regression via applying the least-squares method to regress y_i on $\mathbf{x}_i, i \in \mathcal{N}$. This regression fitting itself says little about whether the resulting regression line $y = \hat{m}(\mathbf{x})$ is a good fit to $(y_i, \mathbf{x}_i)$ or not. However, the example above indicates that, for the purpose of estimating the population average of y, the lack of fit may not matter that much, as long as the "residual" $z_I = y_I - \hat{m}(\mathbf{x}_I)$ has little correlation with $W_I \pi_I$, as two functions of the FPI variable I. Indeed, as discussed in Section 3, we can consider including $\hat{\pi}_I$ in the regression model $\hat{m}(\mathbf{x}_I, \hat{\pi}_I)$. How effective this strategy is in general is a topic of further research.

6. Counterbalancing sub-sampling

6.1 The devastating impact of data defect on effective sample size

A key finding, which has surprised many, from studying the data quality issue is how small the size of our "big data" is when we take into account the data defect. To prove this mathematically, we can equate the mean-squared error (MSE) of $\bar{G}_W$ in (2.1), with the MSE of a simple random sampling estimator of size n_{eff}. This yields (see Meng (2018) for derivation):

$$n_{\text{eff}} \approx \frac{f_W}{1-f_W}\frac{1}{\mathrm{E}[\rho^2_{\tilde{R},G}]} \approx \frac{f_W}{1-f_W}\frac{1}{\rho^2_{\tilde{R},G}}, \tag{6.1}$$

where $f_W = n_W/N$ and the expectation E is with respect to the conditional distribution of $\tilde{R}$ given n_W. It is worthwhile to note that this (conditional) distribution can involve all three types of probability discussed in Section 1.2 because the variations in $\tilde{R}$ can come from multiple sources. For example, in typical opinion surveys, there will be (1) design probability in the sampling indicator, (2) divine probability in formulating the non-response mechanism, and (3) device probability for estimating the mechanism and the weights.

Expression (6.1) is the weighted version/extension of the expression given in Meng (2018) with equal weights, which reveals the devastating impact of a seemingly tiny *ddc*. Suppose our sample is 1% of the population, and it suffers from a half-percent *ddc*. Applying (6.1) (with equal weights) with $f_W = 0.01$ and $\rho_{\tilde{R},G} = 0.005$ yields $n_{\text{eff}} \approx 404$ *regardless of the sample size* n_R. In the case of the 2020 US presidential election, 1% of the voting population is about 1.55 million people, and hence the loss of sample size due to a half percent *ddc* is about 1 - (404 / 1,550,000) > 99.97%. Such seemingly impossible losses have been reported in both election studies (Meng, 2018) and COVID vaccination studies (Bradley et al., 2021). A most devastating consequence of such losses is the "big data paradox": the larger the (apparent) data size, the surer we fool ourselves because our false confidence (in both technical and literal sense) goes up with the erroneous data size, while the actual coverage probability of the incorrectly constructed confidence intervals become vanishingly small (Meng, 2018; Msaouel, 2022).

A positive implication from this revelation, however, is that we can trade much data quantity for data quality, and still end up having statistically more accurate estimates. Of course, in order to reduce the bias, we will need some information about it. If we have reliable information on the value of *ddc*, we can directly adjust for the bias in estimating the population average corresponding to the *ddc*, for example by a Bayesian approach, similar to that taken by Isakov and Kuriwaki (2020) in their scenario analysis. Furthermore, if we have sufficient information to construct reliable weights, we can use the weights to adjust for selection biases as commonly done. Nevertheless, even in such cases, it may still be useful to create a representative miniature of the population out of a biased sample for general purposes, which for example can eliminate many practitioners' anxiety and potential mistakes for not knowing how to properly use the weights. Indeed, few really know how to deal with weights, because "Survey weighting is a mess" (Gelman, 2007).

However, creating a representative miniature out of a biased sample in general is a challenging task, especially because *ddc* can (and will) vary with the variable of interest. Nevertheless, just as weighting is popular tool despite it being far from perfect, let us explore representative miniaturization and see how far we can push the idea. The following example therefore is purely for brainstorming purposes, by looking into a common but challenging scenario, where we have reasonable information or understanding on the direction of the bias, that is, the sign of the *ddc*, but rather vague information about its magnitude. A good example is non-representativeness of election polls because voters tend to not want to disclose their

preferences when they plan to vote for a socially unpopular candidate; we therefore know the direction of the bias, but not much about its degree other than some rough guesses (e.g., a range of 10 percentage points).

6.2 Creating a less biased sub-sample

The basic idea is to use such partial information about the selection bias to design a *biased* sub-sampling scheme to *counterbalance* the bias in the original sample, such that the resulting sub-samples have a *high likelihood* to be less biased than the original sample from our target population. That is, we create a sub-sampling indicator S_I, such that with high likelihood, the correlation between $S_I R_I$ and G_I is reduced, compared to the original $\rho_{R,G}$, to such a degree that it will compensate for the loss of sample size and hence reduce the MSE of our estimator (e.g., the sample average). We say with *high likelihood*, in its non-technical meaning, because without full information on the response/recording mechanism, we can never guarantee such a counterbalance sub-sampling (CBS) would always do better. However, with judicious execution, we can reduce the likelihood of making serious mistakes.

To illustrate, consider the case where y is binary. Let $\Delta = r_1 - r_0$, where r_y is the propensity of responding/reporting for individuals whose responses will take value y: $r_y = \Pr_I(R_I = 1 \,|\, y_I = y)$. If the sample is representative, then like $\rho_{R,G}$, Δ is miniaturized, meaning that it is on the order of $N^{-1/2}$. This is most clearly seen via the easily verifiable identity (see (4.1) of Meng, 2018)

$$\Delta = \frac{\mathrm{Cov}_I(y_I, R_I)}{p(1-p)} = \rho_{R,y}\sqrt{\frac{f_R(1-f_R)}{p(1-p)}}, \tag{6.2}$$

where $p = \Pr_I(y_I = 1)$ and $f_R = \Pr_I(R_I = 1)$, which is the original sampling rate. A key ingredient of CBS is to determine $s_y = P_I(S_I = 1 \,|\, y_I = y, R_I = 1)$ for $y = 0, 1$, that is, the sub-sampling probabilities of individuals who reported $y = 1$ and $y = 0$, respectively.

To determine the beneficial choices, let $f_S = \Pr_I(S_I = 1 \,|\, R_I = 1)$ be the sub-sampling rate, and $\Delta_S = s_1 r_1 - s_0 r_0$. Then by applying (2.2) (with equal weights) and (6.2) to both the sample average and the sub-sample average, we see that the sub-sample average has smaller (actual) error in magnitude if and only if

$$\left(\frac{\Delta_S}{f_S f_R}\right)^2 < \left(\frac{\Delta}{f_R}\right)^2 \Leftrightarrow f_S^2 > \left(\frac{\Delta_S}{\Delta}\right)^2. \tag{6.3}$$

Writing $r = r_1/r_0$ and $s = s_1/s_0$, the right-hand side of (6.3) becomes

$$\left[sp^* + (1-p^*)\right]^2 > \left(\frac{rs-1}{r-1}\right)^2, \tag{6.4}$$

where $p^* = \Pr_I(y_I = 1 \mid R_I = 1)$ is observed in the original sample, which should remind us that p^* may be rather different from the p we seek, because of the biased R-mechanism.

An immediate choice to satisfy (6.4) is to set $s = r^{-1}$, which of course typically is unrealistic because if we know the value of r, then the problem would be a lot simpler. To explore how much leeway we have in deviating from this ideal choice, let $\delta = r - 1$, we can then show that (6.4) is equivalent to

$$(s-1)\left\{[1+(1+p^*)\,\delta]\,(s-1)+2\delta\right\} < 0. \tag{6.5}$$

This tells precisely the permissible choices of s without over-correcting (in the magnitude of the resulting bias):

(i) When $r > 1$, i.e., $\delta > 0$, we can take any s such that

$$\frac{[1-(1-p^*)\,\delta]_+}{1+(1+p^*)\,\delta} \le s < 1; \tag{6.6}$$

(ii) When $r < 1$, i.e., $\delta < 0$, we can take any s such that

$$1 < s \le \frac{1-(1-p^*)\,\delta}{[1+(1+p^*)\,\delta]_+}. \tag{6.7}$$

This pair of results confirms a number of our intuitions, but also offers some qualifications that are not so obvious. Since we sub-sample to compensate for the bias in the original sample, s and r must stay on the opposite side of 1, i.e., $(s-1)(r-1) = (s-1)\,\delta < 0$, as seen in (6.6)-(6.7). To prevent over corrections, some limits are needed, but it is also possible that the initial bias is so bad that no sub-sampling scheme can make things worse, which is reflected by the positivizing function $[x]_+$ in the two expressions above. However, the expressions for the limits as well as for the thresholds to activate the positivizing functions are not so obvious. Nor is it obvious that these expressions depend on the unknown p indirectly via the observed p^*, and hence only prior knowledge of r is required for implementing or assessing CBS.

This observation suggests that it is possible to implement a beneficial CBS when we can borrow information from other surveys (or studies) where the response/recording behaviors are of similar nature. For example, we may learn that a previous similar survey had $r = 1.5$ (e.g., those with $y = 1$ had 6% of chance to be recorded, and those with $y = 0$ had only 4% chance). Taking into account the uncertainty in the similarity between the two surveys, we might feel comfortable to place (1.2, 1.8) as the plausible range for r in the current study. Suppose we observe $p^* = 0.6$, this means that the maximum – over the range $r \in (1.2, 1.8)$ – of the lower bound on the permissible s as given in (6.6) is

$$\frac{[1-(1-0.6)\,(r-1)]_+}{1+1.6\,(r-1)} = \frac{[1.4-0.4r]_+}{1.6r-0.6} \le \frac{1.4-0.4\times 1.2}{1.6\times 1.2-0.6} = 0.7. \tag{6.8}$$

Therefore, as long as we choose $s \in [0.7, 1)$, we are unlikely to over-correct. The price we pay for this robustness is that the resulting sub-sample is not as good quality as it can be, for example, when the underlying r for the current study is indeed 1.5 (in expectation). Choosing any $s \in [0.7, 1)$ will not provide the full correction as provided by $s = 1/r = 0.67$, that is, the sub-sample average will still have a positive bias but with a smaller MSE compared to the original sample average. Of course both the feasibility and effectiveness of such CBS need to be carefully investigated before it can be recommended for general consumption, especially going beyond binary y. The literature on inverse sampling (Hinkins, Oh and Scheuren, 1997; Rao, Scott and Benhin, 2003) is of great relevance for such investigations, because it also aims to produce simple random samples via subsampling, albeit with a different motivation (to turn complex surveys into simple ones for ease of analysis).

7. Probability sampling as aspiration, not prescription

As it should be clear from the definition of *ddc*, it is not directly estimable from the biased sample alone. One therefore naturally would (and should) question how useful *ddc* is or could be. The answer turns out to be an increasingly long one thanks to *ddc* being model-free and hence a versatile data quality metric for both probability samples and non-probability samples. Its usefulness for generating theoretical insights is demonstrated by its role in quantifying the data quality-quantify trade-off via effective sample size as seen in (6.1), in understanding simulation errors in quasi-Monte Carlo as explored in Hickernell (2016), and in anticipating the "double-plus robustness" phenomenon as presented in Section 5. Its methodological usages are illustrated by the scenario analyses for the 2020 US Presidential election (Isakov and Kuriwaki, 2020) and for the COVID-19 vaccination assessments (Bradley et al., 2021). Its practical implications can be found in epidemiological studies (Dempsey, 2020), particle physics (Courtoy, Houston, Nadolsky, Xie, Yan and Yuan, 2022), and political polling (Bailey, 2023).

Not surprisingly, these practical applications found the notion of *ddc* and the underlying error decomposition (2.2) helpful because of the non-probability samples they need to deal with, either due to distortions to the probability samples such as by a biased non-response mechanism or due to selection biases in the first place such as selective COVID-19 testing. Professor Wu's overview, and the many references cited there and in this discussion, should make it clear that non-probability samples are *almost surely* everywhere. I am invoking this strong probabilistic phrase not merely for its humorous value. When we consider the unaccountably many possible values for the mean of *ddc*, the probability – however we construct it to capture the wild west of data collection processes out there – that it will land precisely on zero must be zero. This zero mean is a necessary condition for the sample to be a probability sample, because a probability sample implies that *ddc* must be of the order of $N^{-1/2}$ order (Meng, 2018), which is impossible when its mean is non-zero (asymptotically). This observation suggests that we should move away from our tradition of treating probability sampling as a centerpiece and then try to model the much larger world of non-probability samples as "deviations" from it. Instead, we should start with studying samples with general collection mechanisms using tools or concepts such as *ddc*, and then treat (design)

probability samples as the very special, ideal case – always an aspiration, but never the only prescription for action.

Acknowledgements

I am grateful to Editor Jean-François Beaumont for inviting me to discuss Changbao Wu's timely and thought-provoking overview. I thank James Bailie, Radu Craiu, Adel Daoud, Andrew Gelman, Stas Kolenikov, Rod Little, Cory McCartan, Kelly McConville, James Robins, Zhiqiang Tan, and Li-Chun Zhang for moral endorsement and for constructive criticisms. I also thank NSF for partial financial support, and Steve Finch for careful proofreading.

References

Bailey, M.A. (2023). *Polling at a Crossroads – Rethinking Modern Survey Research*. Cambridge University Press.

Beaumont, J.-F., and Rao, J.N.K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *Survey Statistician*, 83, 11-22.

Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518), 859-877.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, Z.-L. and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), 695-700.

Buelens, B., Burger, J. and van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *International Statistical Review*, 86(2), 322-343.

Courtoy, A., Houston, J., Nadolsky, P., Xie, K., Yan, M. and Yuan, C.-P. (2022). Parton distributions need representative sampling. *arXiv preprint arXiv:2205.10444.*

Craiu, R.V., Gong, R. and Meng, X.-L. (2022). Six statistical senses. *arXiv preprint arXiv:2204.05313.*

David Peat, F. (2002). *From Certainty to Uncertainty: The Story of Science and Ideas in the Twentieth Century.* Joseph Henry Press.

Dempsey, W. (2020). The hypothesis of testing: Paradoxes arising out of reported coronavirus case-counts. *arXiv preprint arXiv:2005.10425*.

Dwork, C. (2008). Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, Springer, 1-19.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153-164.

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*, (Special Issue 2), June 24, 2022. https://hdsr.mitpress.mit.edu/pub/ld4smnnf.

Gong, R., Groshen, E.L. and Vadhan, S. (2022). Harnessing the known unknowns: Differential privacy and the 2020 Census. *Harvard Data Science Review*, (Special Issue 2), June 24 2022. https://hdsr.mitpress.mit.edu/pub/fgyf5cne.

Han, P., and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100(2), 417-430.

Hartley, H.O., and Ross, A. (1954). Unbiased ratio estimators. *Nature*, 174(4423), 270-271.

Hickernell, F.J. (2016). The trio identity for Quasi-Monte Carlo error. In International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer, 3-27.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-21. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3101-eng.pdf.

Isakov, M., and Kuriwaki, S. (2020). Towards principled unskewing: Viewing 2020 election polls through a corrective Lens from 2016. *Harvard Data Science Review*, 2(4), Nov. 3, 2020. https://hdsr.mitpress.mit.edu/pub/cnxbwum6.

Kang, J.D.Y., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4), 523-539.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Li, X., and Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-*p*-zero-*n*: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association*, 116(533), 353-367.

Little, R., and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14(3), 949-968.

Liu, Y., Gelman, A. and Chen, Q. (2021). Inference from non-random samples using Bayesian machine learning. *arXiv preprint arXiv:2104.05192*.

Lo, A.W. (2017). Adaptive markets. In *Adaptive Markets*. Princeton University Press.

Lohr, S., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475), 1019-1030.

Lohr, S.L. (2021). *Sampling: Design and Analysis*. Chapman and Hall/CRC.

Luque-Fernandez, M.A., Schomaker, M., Rachet, B. and Schnitzer, M.E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, 37(16), 2530-2546.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science*, (Eds., Lin et al.), CRC Press.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (i) Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, 12(2), 685-726.

Meng, X.-L. (2021). Enhancing (publications on) data quality: Deeper data minding and fuller data confession. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(4), 1161-1175.

Msaouel, P. (2022). The big data paradox in clinical practice. *Cancer Investigation*, 1-27.

Pfeffermann, D. (2017). Bayes-based non-bayesian inference on finite populations from non-representative samples: A unified approach. *Calcutta Statistical Association Bulletin*, 69(1), 35-63.

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling. *Survey Methodology*, 29, 2, 107-128. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6787-eng.pdf.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, Indianapolis, IN, 1999, 6-10.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427), 846-866.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussions). *Journal of the American Statistical Association*, 94(448), 1096-1146.

Slavkovic, A., and Seeman, J. (2022). Statistical data privacy: A song of privacy and utility. *arXiv preprint arXiv:2205.03336*.

Tan, Y.V., Flannagan, C.A.C. and Elliott, M.R. (2019). "Robust-Squared" imputation models using Bart. *Journal of Survey Statistics and Methodology*, 7(4), 465-497.

Tan, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science*, 22(4), 560-568.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3), 661-682.

Tan, Z. (2013). Simple design-efficient calibration estimators for rejective and high-entropy sampling. *Biometrika*, 100(2), 399-415.

Van Buuren, S., and Oudshoorn, K. (1999). *Flexible Multivariate Imputation by MICE*. Leiden: TNO.

van der Laan, M.J., and Gruber, S. (2009). Collaborative double robust targeted penalized maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, 246.

van der Laan, M.J., and Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 6(1).

van der Laan, M.J., and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), 980-991.

Wu, C. (2022). Statistical inference with non-probability survey samples (with discussions). *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

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

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 445-465.

Zhang, G., and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3), 911-918.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2), 103-113.

Survey Methodology, December 2022
Vol. 48, No. 2, pp. 361-366
Statistics Canada, Catalogue No. 12-001-X

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

Zhonglei Wang and Jae Kwang Kim[1]

Abstract

Statistical inference with non-probability survey samples is a notoriously challenging problem in statistics. We introduce two new methods of nonparametric propensity score technique for weighting in the non-probability samples. One is the information projection approach and the other is the uniform calibration in the reproducing kernel Hilbert space.

Key Words: Information projection; Uniform function calibration; Data integration.

1. Introduction

We would like to congratulate Dr. Changbao Wu on the outstanding work in non-probability sampling. Even though probability sampling served as a golden standard tool for finite population inference in the past decades, it has recently become tarnished gold due to low response rates and high costs. Non-probability sampling, on the other hand, is popular due to its feasibility and low cost (Couper, 2000; Kaplowitz, Hadlock and Levine, 2004). More importantly, non-probability sampling, such as a web survey, can quickly gather up-to-date information when compared to a probability sample. However, because the selection mechanism is unavailable for non-probability sampling, failing to correct the selection bias in analyzing a non-probability sample may result in inefficiency or even erroneous inference. As a result, adjusting the selection bias for a non-probability sample is a fundamental topic for survey sampling researchers, and this work presents the most comprehensive answers to this subject.

Dr. Wu's research, in particular, includes a thorough examination of propensity score (PS) techniques. Those PS techniques, on the other hand, have drawbacks. First, even for a correctly specified PS model, the inverse probability weighting estimator may be inefficient due to small estimated propensity scores. One alternative is post-stratification, as stated in Section 5 of the paper, although there is no clear guidance on how to choose K. Furthermore, in practice, correctly specifying a PS model is difficult. While doubly robust estimation can help to safeguard a bad PS model, the final estimator is problematic when both the PS and regression models are incorrect (Kang and Schafer, 2007).

To overcome the misspecification of the PS model, Dr. Wu has mentioned several nonparametric methods, including a kernel method and a tree-based method. In this discussion, we would like to expand on this direction and provide two more methods to augment the study. One is based on a density ratio model using information projection (Csiszár and Shields, 2004), and the other is by uniformly calibrating functions over a reproducing kernel Hilbert space (RKHS). As explained by Wahba (1990), RKHS is a very flexible function space for approximation. Instead of estimating the propensity scores, we aim at

1. Zhonglei Wang, Wang Yanan Institute for Studies in Economics and School of Economics, Xiamen University, Xiamen, Fujian, People's Republic of China; Jae Kwang Kim, Iowa State University, Ames, IA 50011, USA. E-mail: jkim@iastate.edu.

estimating the sampling weights $\{(\pi_i^A)^{-1} : i \in S_A\}$ to avoid possible inefficiency due to small estimated propensity scores.

Denote S_A and S_B to be the index sets for the non-probability and reference probability samples, respectively, and the corresponding sample sizes are n_A and n_B. Let $\{(y_i, \mathbf{x}_i) : i \in S_A\}$ and $\{(\mathbf{x}_i, d_i^B) : i \in S_B\}$ be available, where y_i and $\mathbf{x}_i$ are the study variable and auxiliary vector for the i^{th} unit and d_i^B is the design weight for $i \in S_B$.

The paper is organized as follows. In Section 2, we introduce the information projection approach. In Section 3, we introduce the basic idea of uniform calibration. Some concluding remarks are made in Section 4.

2. Information projection approach

Suppose that we are interested in estimating parameter $\boldsymbol{\theta}_0$ defined through $E_N\{U(\boldsymbol{\theta}; \mathbf{X}, Y)\} = 0$, where $E_N(\cdot)$ is the expectation with respect to the population empirical distribution $\Pr\{(\mathbf{X}, Y) = (\mathbf{x}_i, y_i)\} = N^{-1}$ for $i = 1, \ldots, N$ and 0 otherwise, and $U(\boldsymbol{\theta}; \mathbf{x}, y)$ is a certain estimating function. For example, $U(\theta; \mathbf{x}, y) = y - \theta$ corresponds to $\mu_y = N^{-1}\sum_{i=1}^N y_i$ in the paper. We wish to obtain an estimator of $(\pi_i^A)^{-1}$, $\pi_i^A = \Pr(R_i = 1 \mid \mathbf{x}_i, y_i)$, and $R_i = 1$ if $i \in S_A$ and 0 otherwise.

To estimate $\{(\pi_i^A)^{-1} : i \in S_A\}$, we may use the relationship in the density ratio function. First, we consider a super-population model ξ, and let $f_0(\mathbf{x}, y)$ and $f_1(\mathbf{x}, y)$ be the density functions of $(\mathbf{x}, y)$ given $R = 0$ and $R = 1$, respectively. Denote the density ratio function to be

$$r(\mathbf{x}, y) = \frac{f_0(\mathbf{x}, y)}{f_1(\mathbf{x}, y)},$$

and by the Bayes formula, we have

$$(\pi_i^A)^{-1} = 1 + \frac{\Pr(R_i = 0)}{\Pr(R_i = 1)} r(\mathbf{x}_i, y_i). \tag{2.1}$$

Thus, there is a one-to-one relationship between $(\pi_i^A)^{-1}$ and $r(\mathbf{x}_i, y_i)$.

Under assumption A1, we can show that $r(\mathbf{x}, y) = r(\mathbf{x})$. In this section, we make a more general assumption that there exists $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_L(\mathbf{x}))^{\mathrm{T}}$ such that

$$R \perp Y \mid \mathbf{b}(\mathbf{x}). \tag{2.2}$$

Rosenbaum and Rubin (1983) called $\mathbf{b}(\mathbf{x})$ in (2.2) balancing scores.

To estimate the density ratio function $r(\mathbf{x})$, we minimize the Kullback-Leibler divergence

$$Q(f_0) = \int \log(f_0 / f_1) f_0 \, \mathrm{d}\mu \tag{2.3}$$

with respect to f_0 subject to some constraint, where both f_0 and f_1 are absolutely continuous with respect to a σ-finite measure μ. Regarding the constraint, we may use the following one

$$\Pr(R_i=1)\int \mathbf{b}(\mathbf{x})f_1(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) + \Pr(R_i=0)\int \mathbf{b}(\mathbf{x})f_0(\mathbf{x})\mu(\mathrm{d}\mathbf{x}) = E_\xi\{\mathbf{b}(\mathbf{X})\}, \tag{2.4}$$

where $E_\xi(\cdot)$ is the expectation with respect to the super-population model ξ. That is, given $f_1(\mathbf{x})$, we can find $f_0(\mathbf{x})$ to minimize (2.3) under a calibration constraint with respect to $\mathbf{b}(\mathbf{x})$.

By Lemma 3.1 of Wang and Kim (2021), the optimized conditional density function satisfies

$$f_0^*(\mathbf{x}) = f_1(\mathbf{x})\frac{\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{b}(\mathbf{x})\}}{E_1\left[\exp\{\boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{b}(\mathbf{x})\}\right]}, \tag{2.5}$$

where $\boldsymbol{\lambda}_1$ is chosen to satisfy (2.4). Note that the solution (2.5) is equivalent to

$$\log\{r(\mathbf{x};\boldsymbol{\lambda})\} = \lambda_0 + \boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{b}(\mathbf{x}) \tag{2.6}$$

for the density ratio function $r(\mathbf{x})$, where $\boldsymbol{\lambda}=(\lambda_0,\boldsymbol{\lambda}_1^{\mathrm{T}})^{\mathrm{T}}$, and λ_0 is a normalizing constant satisfying $\int r(\mathbf{x};\boldsymbol{\lambda})f_1(\mathbf{x})\mu(\mathrm{d}\mathbf{x})=1$. Thus, the information projection finds the best model for propensity score function.

Once the model is determined as in (2.6), we need to estimate the model parameters. Because of the moment constraints in (2.4), the sample-version estimating equation for $\boldsymbol{\lambda}$ is the calibration equation given by

$$\frac{n_A}{N}\sum_{i=1}^{N} R_i[1,\mathbf{b}(\mathbf{x}_i)]\left[1+\frac{1-n_A}{n_A}\exp\{\lambda_0+\boldsymbol{\lambda}_1^{\mathrm{T}}\mathbf{b}(\mathbf{x}_i)\}\right] = \left[1,\frac{1}{N}\sum_{i\in S_B} d_i^B\mathbf{b}(\mathbf{x}_i)\right]. \tag{2.7}$$

Here, since $E_\xi\{\mathbf{b}(\mathbf{X})\}$ is not available, we use its estimate $N^{-1}\sum_{i\in S_B} d_i^B\mathbf{b}(\mathbf{x}_i)$. Once the parameter estimate $\hat{\boldsymbol{\lambda}}$ is obtained, we can construct

$$\hat{\omega}_i = 1 + \frac{1-n_A}{n_A}\exp\{\hat{\lambda}_0+\hat{\boldsymbol{\lambda}}_1^{\mathrm{T}}\mathbf{b}(\mathbf{x}_i)\}$$

as the final PS weights. The parameter of interest can be estimated by solving $N^{-1}\sum_{i\in S_A}\hat{\omega}_i U(\boldsymbol{\theta};\mathbf{x}_i,y_i)=0$ for $\boldsymbol{\theta}$.

Wang and Kim (2021) developed this framework under the non-probability sampling setup where $\mathbf{x}_i$ are available throughout the finite population. Consistency and the asymptotic normality can be developed under the assumption that $E\{U(\boldsymbol{\theta};\mathbf{x},Y)\,|\,\mathbf{x}\}$ lies in the linear space generated by $\{b_1(\mathbf{x}),\ldots,b_L(\mathbf{x})\}$. Instead of assuming the availability of $\{\mathbf{x}_i: i=1,\ldots,N\}$ as in Wang and Kim (2021), there only exists a reference probability sample $\{(\mathbf{x}_i,d_i^B):\ i\in S_B\}$. If the probability sample S_B is a census, then the method above reduces to the one considered by Wang and Kim (2021), except that we consider a finite population parameter $\boldsymbol{\theta}_0$. In Section 11.2 of Kim and Shao (2021), the information projection approach is called the

maximum entropy method and applied to the data integration problem. In the simulation study presented in example 11.1 of the book, the proposed information projection method shows better performance than the methods of Chen, Li and Wu (2020) and Elliott and Valliant (2017).

3. Uniform calibration approach

Calibration is commonly used to improve the representativeness of a non-probability sample, but existing methods, including the information projection approach mentioned in Section 2, are based on calibrating a set of pre-specified functions. However, it is hard to correctly specify them for calibration in practice. In this section, we propose a general framework for uniformly calibrating functions in an RKHS. Instead of considering a parametric form for $E_\xi(Y \mid \mathbf{x})$ in (3.1), we only assume $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i)$, where $m(\mathbf{x})$ is a smooth function satisfying certain conditions.

We still consider (2.1) under the assumption A1. Instead of assuming a set of pre-specified functions $\mathbf{b}(\mathbf{x})$, we propose to estimate $\{r_i : i \in S_A\}$ by the following optimization,

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \geq 0}{\operatorname{argmin}} \left[\sup_{u \in H} \left\{ \frac{S(\boldsymbol{\gamma}, u)}{\|u\|_2^2} - \lambda_1 \frac{\|u\|_H^2}{\|u\|_2^2} \right\} + \lambda_2 Q_A(\boldsymbol{\gamma}) \right], \tag{3.1}$$

where $\boldsymbol{\gamma} = (r_1, \ldots, r_N)$, $r_i = 0$ for $i \notin S_A$, $\boldsymbol{\gamma} \geq 0$ is equivalent to $r_i \geq 0$ for $i = 1, \ldots, N$, H is an RKHS,

$$S(\boldsymbol{\gamma}, u) = \left[N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) r_i \right\} u(\mathbf{x}_i) - N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i) \right]^2, \tag{3.2}$$

$\|u\|_2^2 = (n_A + n_B)^{-1} \sum_{i \in S_A \cup S_B} u(\mathbf{x}_i)^2$, $\|u\|_H$ is the norm associated with the RKHS, $Q_A(\boldsymbol{\gamma})$ is a general penalty on $\boldsymbol{\gamma}$ to avoid overfitting, and λ_1 and λ_2 are two tuning parameters; see Wahba (1990) for a detailed introduction about the RKHS.

The intuition for the optimization (3.1) is briefly discussed. First, if r_i approximates the true density ratio $r(\mathbf{x}_i)$ well, the bias of the first term in (3.1) is negligible for estimating $N^{-1} \sum_{i=1}^N u(\mathbf{x}_i)$ for $u \in H$. Besides, $N^{-1} \sum_{i \in S_B} d_i^B u(\mathbf{x}_i)$ is design-unbiased. Thus, $S(\boldsymbol{\gamma}, u)$ balances two estimators for $N^{-1} \sum_{i=1}^N u(\mathbf{x}_i)$, and it is small if r_i approximately equals $r(\mathbf{x}_i)$ for $i \in S_A$. However, $S(\boldsymbol{\gamma}, u)$ is not scale invariant, and we have $S(\boldsymbol{\gamma}, cu) = c^2 S(\boldsymbol{\gamma}, u)$ for $c \in \mathbb{R}$. Thus, we use $\|u\|_2^2$ to make it scale-invariant. The term $\lambda_1 \|u\|_H^2$ is used to penalize the smoothness of the function u for $u \in H$. There exist different choices for $Q_A(\boldsymbol{\gamma})$. For example, $Q_A(\boldsymbol{\gamma}) = \sum_{i \in S_A} \{1 + (N n_A^{-1} - 1) r_i\}^2$ corresponds to penalizing extreme values for the sampling weights, and Wong and Chan (2018) investigated a similar problem assuming the availability of $\{\mathbf{x}_i : i = 1, \ldots, N\}$. The optimization (3.1) can be viewed as a "minmax" problem, and if $m \in H$, the estimated density ratios $\{\hat{r}_i : i \in S_A\}$ may lead to a reasonably good estimator

$$\hat{\mu}_{uc} = N^{-1} \sum_{i \in S_A} \left\{ 1 + \left(\frac{N}{n_A} - 1 \right) \hat{r}_i \right\} y_i. \tag{3.3}$$

Uniform calibration is a new method for non-probability sampling, and there are some technical challenges in (3.1). For example, how to incorporate the design properties of S_B when establishing the theoretical properties of (3.3) has not be fully investigated, and we have finished a working paper about this topic (Wang, Mao and Kim, 2022). The kernel-based method is computationally expensive, especially when the sample sizes are large. It may be interesting to propose a more computationally efficient algorithm for the uniform calibration problem. One possible answer is to consider some other functional spaces, such as the one spanned by B-splines. In addition, it is also of interest to consider how to incorporate more than one reference probability sample, and how to formulate a uniform calibration if we have different covariates in different reference probability samples.

4. Concluding remarks

Propensity score weighting is an important tool for correcting selection bias in the nonprobability sampling. Dr. Changbao Wu made significant contributions on this important topic. In addition to the two additional methods, the empirical likelihood (EL) approach of Qin, Leung and Shao (2002) is potentially useful as another tool for propensity score weighting. In particular, the EL-based weighting method is applicable even under informative sampling. Further investigation on this direction will be explored elsewhere.

References

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly*, 64(4), 464-494.

Csiszár, I., and Shields, P.C. (2004). *Information Theory and Statistics: A Tutorial*.

Elliott, M., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264.

Kang, J.D., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22(4), 523-539.

Kaplowitz, M.D., Hadlock, T.D. and Levine, R. (2004). A comparison of Web and mail survey response rates. *Public Opinion Quarterly*, 68(1), 94-101.

Kim, J.K., and Shao, J. (2021). *Statistical Methods for Handling Incomplete Data*, second edition. CRC press.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under non-ignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wang, H., and Kim, J.K. (2021). Information projection approach to propensity score estimation for handling selection bias under missing at random. *arXiv:2104.13469*, 1-34.

Wang, Z., Mao, X. and Kim, J.K. (2022). Functional calibration under non-probability survey sampling. Submitted (https://arxiv.org/abs/2204.09193).

Wong, R.K., and Chan, K.C.G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1), 199-213.

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.