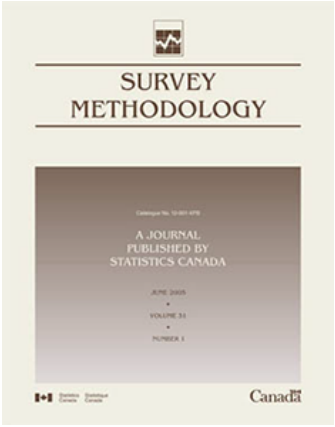
Catalogue no. 12-001-X
ISSN 1492-0921

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.