

IOWA STATE UNIVERSITY

Digital Repository

Statistics Preprints

Statistics

1-29-2018

Sampling techniques for big data analysis in finite population inference

Jae Kwang Kim

Iowa State University, jkim@iastate.edu

Zhonglei Wang

Iowa State University, wangzl@iastate.edu

Follow this and additional works at: https://lib.dr.iastate.edu/stat_las_preprints

 Part of the [Design of Experiments and Sample Surveys Commons](#), [Probability Commons](#), and the [Statistical Methodology Commons](#)

Recommended Citation

Kim, Jae Kwang and Wang, Zhonglei, "Sampling techniques for big data analysis in finite population inference" (2018). *Statistics Preprints*. 136.

https://lib.dr.iastate.edu/stat_las_preprints/136

This Article is brought to you for free and open access by the Statistics at Iowa State University Digital Repository. It has been accepted for inclusion in Statistics Preprints by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

Sampling techniques for big data analysis in finite population inference

Abstract

In analyzing big data for finite population inference, it is critical to adjust for the selection bias in the big data. In this paper, we propose two methods of reducing the selection bias associated with the big data sample. The first method uses a version of inverse sampling by incorporating auxiliary information from external sources, and the second one borrows the idea of data integration by combining the big data sample with an independent probability sample. Two simulation studies show that the proposed methods are unbiased and have better coverage rates than their alternatives. In addition, the proposed methods are easy to implement in practice.

Keywords

Data integration, inverse sampling, non-probability sample, selection bias

Disciplines

Design of Experiments and Sample Surveys | Probability | Statistical Methodology

Comments

This is a manuscript that has been accepted for publication in *International Statistical Review*:

<https://arxiv.org/abs/1801.09728>.

Sampling techniques for big data analysis in finite population inference

Jae Kwang Kim and Zhonglei Wang

Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.

E-mails: jkim@iastate.edu and wangzl@iastate.edu

Summary

In analyzing big data for finite population inference, it is critical to adjust for the selection bias in the big data. In this paper, we propose two methods of reducing the selection bias associated with the big data sample. The first method uses a version of inverse sampling by incorporating auxiliary information from external sources, and the second one borrows the idea of data integration by combining the big data sample with an independent probability sample. Two simulation studies show that the proposed methods are unbiased and have better coverage rates than their alternatives. In addition, the proposed methods are easy to implement in practice.

Key words: Data integration; inverse sampling; non-probability sample; selection bias.

1 Introduction

Probability sampling is a scientific tool for obtaining a representative sample from a target finite population. Formally, a probability sample has the property that every element in the finite population has a known and nonzero probability of being selected. Probability sampling can be used to construct valid statistical inferences for finite population parameters. Survey sampling is an area of statistics that deals with constructing efficient probability sampling designs and corresponding estimators. Classical approaches in survey sampling are discussed in Cochran (1977), Särndal *et al.* (1992) and Fuller (2009).

Despite the merits of probability samples, Baker *et al.* (2013) argue that it becomes common to get non-probability samples, which may not represent the target population properly. Besides, collecting a strict probability sample is almost impossible in certain areas due to unavoidable issues such as frame undercoverage and nonresponse. The increasing prevalence of non-probability samples, such as web panels, makes methods for non-probability samples even more important. Keiding and Louis (2016) address the challenges in using non-probability samples for making inferences. Elliott and Valliant (2017) review the weighting methods for reducing the selection bias in non-probability samples. Rivers (2007) proposes nearest neighbor imputation matching for combining information from survey data and big data. Bethlehem (2016) discusses sample matching methods for handling non-probability samples.

Big data is one example of such non-probability sample. The Four Vs (volume, velocity, variety and veracity) of big data and its implication to statistical inference are nicely discussed in Franke *et al.* (2016). While use of big data for predictive analysis is a hot area of research (Efron and Hastie, 2016), its use for finite population inference is not well investigated in the literature. Tam (2015) discusses a statistical framework for analyzing big data for official statistics, particularly in agricultural statistics. Rao and Molina (2015) discuss using the area-level summary of big data as one of the covariates in the linking model for small area estimation. Tam and Kim (2018) cover some ethical challenges of big data for official statisticians and discuss some preliminary methods of correcting for selection bias in big data.

One of the benefits of using big data is, as pointed out by Tam and Clarke (2015), in the cost effectiveness in the production of official statistics. However, there are still great challenges when using big data for finite population inference. The most critical issue is how to handle selection bias in the big data sample (Meng, 2018). Adjusting for the selection bias in big data is an important practical problem in survey sampling.

In this paper, we discuss how some of the sampling techniques can be applied in harnessing big data for finite population inference. By treating the selection bias in the big data sample as a missing data problem, we propose two approach of handling big data

in survey sampling. The first approach is based on inverse sampling, which is a special case of two-phase sampling, and a novel inverse sampling method is proposed to obtain a representative sample from the big data. The second approach is based on the weighting method using the auxiliary information obtained from another independent probability sample. Combining information from two data sources, often called data integration, is also a hot area of research in survey sampling. In the proposed method, an independent probability sample is used to estimate the parameters of the propensity score model for the big data sample.

The paper is organized as follows. In Section 2, the basic setup is introduced and the selection bias of big data is discussed. In Section 3, an inverse sampling method is proposed. In Section 4, a propensity score weighting approach using data integration is discussed. Results from two limited simulation studies are presented in Section 5. Some concluding remarks are made in Section 6.

2 Basic Setup

Consider a finite population $\{y_i : i \in U\}$, where y_i is the i -th observation of the study variable Y , and $U = \{1, \dots, N\}$ is the corresponding index set with known size N . A big data sample $\{y_i : i \in B\}$ is available with $B \subset U$. Specifically, $\delta_i = 1$ if $i \in B$ and $\delta_i = 0$ otherwise, and assume that y_i is observed only when $\delta_i = 1$. We are interested in estimating the population mean $\bar{Y}_N = N^{-1} \sum_{i=1}^N y_i$.

From the big data sample B , we can estimate \bar{Y}_N by $\bar{Y}_B = N_B^{-1} \sum_{i=1}^N \delta_i y_i$, where $N_B = \sum_{i=1}^N \delta_i$ is the known size of B . Given $\{\delta_i : i \in U\}$, the error of \bar{Y}_B can be written as

$$\bar{Y}_B - \bar{Y}_N = \frac{1}{f_B} \text{Cov}(\delta, Y)$$

where $f_B = N_B/N$ and

$$\text{Cov}(\delta, Y) = \frac{1}{N} \sum_{i=1}^N (\delta_i - \bar{\delta}_N)(y_i - \bar{Y}_N)$$

with $\bar{\delta}_N = N^{-1} \sum_{i=1}^N \delta_i$. Thus, we have

$$E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = \frac{1}{f_B^2} E_\delta\{\text{Cov}(\delta, Y)^2\}, \quad (1)$$

where $E_\delta(\cdot)$ denotes the expectation with respect to the random mechanism for δ_i .

If the random mechanism for δ_i is based on Bernoulli sampling, where the inclusion indicators follow a Bernoulli distribution with success probability f_B independently, we can obtain

$$\begin{aligned} E_\delta\{\text{Cov}(\delta, Y)^2\} &= [E_\delta\{\text{Cov}(\delta, Y)\}]^2 + \text{Var}_\delta\{\text{Cov}(\delta, Y)\} \\ &= 0 + \frac{1}{N^2} \sum_{i=1}^N (y_i - \bar{Y}_N)^2 f_B(1 - f_B) = \frac{1}{N} f_B(1 - f_B) \sigma^2 \end{aligned}$$

with $\sigma^2 = N^{-1} \sum_{i=1}^N (y_i - \bar{Y}_N)^2$. Thus, under Bernoulli sampling, (1) reduces to

$$E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = \frac{1}{N_B} (1 - f_B) \sigma^2,$$

which is consistent with the classical theory for Bernoulli sampling with sample size $n = N_B$. For general cases, (1) can be expressed as

$$\begin{aligned} E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} &= \frac{1}{f_B^2} E_\delta\{\text{Corr}(\delta, Y)^2 \text{Var}(\delta) \text{Var}(Y)\} \\ &= E_\delta\{\text{Corr}(\delta, Y)^2\} \times \left(\frac{1}{f_B} - 1\right) \times \sigma^2, \end{aligned} \quad (2)$$

where the second equality follows from

$$\text{Var}(\delta) = \frac{1}{N} \sum_{i=1}^N (\delta_i - \bar{\delta}_N)^2 = f_B(1 - f_B).$$

Equality (2) is also presented in Meng (2018). Although there are three terms in (2) determining the selection bias of \bar{Y}_B , the first term, $E_\delta\{\text{Corr}(\delta, Y)^2\}$, is the most critical one. Meng (2018) calls the term *Data Defect Index* (DDI), which determines the level of departure from simple random sampling. Under equal probability sampling designs such that $E_\delta(\delta_i) = f_B$, we have $E_\delta\{\text{Corr}(\delta, Y)\} = 0$ and DDI is of order $O(1/N)$, which implies $E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = O(N_B^{-1})$. For other sampling designs with $E_\delta\{\text{Corr}(\delta, Y)\} \neq 0$, the DDI becomes significant with order $O(1)$, which implies $E_\delta\{(\bar{Y}_B - \bar{Y}_N)^2\} = O(N_B^{-1}N - 1)$.

Therefore, a non-probability sampling design with $E_{\delta}\{\text{Corr}(\delta, Y)\} \neq 0$ makes the analysis results subject to selection bias.

In this paper, we show how to use some of the existing sampling techniques to reduce the selection bias of the big data sample and make the resulting analysis valid. We consider two techniques, one is inverse sampling and the other is survey data integration.

3 Inverse sampling

When the distribution of the study variable for the big data sample differs systematically from that for the target population, the big data sample does not necessarily represent the target population. An important question in this respect is whether we can use auxiliary variables, external to the big data sample, to correct for the selection bias. In this section, we consider a novel inverse sampling approach to address this problem. The proposed inverse sampling can be viewed as a special case of two-phase sampling (e.g., Breidt and Fuller, 1993; Rao and Sitter, 1995; Hidiroglou, 2001; Kim, *et al.* 2006; Stukel and Kott, 1996). The first-phase sample is the big data sample, which is subject to selection bias. The second-phase sample is a subsample of the first-phase sample to correct the selection bias of the big data sample. Inverse sampling is originally proposed as a way of obtaining a simple random sample from a sample obtained from a complex sampling design. For some classical designs, such as stratified sampling, the inverse sampling algorithm is presented by Hinkins *et al.* (1997) and Rao *et al.* (2003). Tillé (2016) applies the inverse sampling concept to a quota sample. We address the application of inverse sampling to big data subject to selection bias.

Unlike the classical two-phase sampling, the first-phase sample in our setup is the big data itself, and we have no control over it. Thus, we first use some external source to determine the level of selection bias in the big data. This step can be called weighting step, as the importance weights are computed for each element in the big data sample. The second step is to select the second phase sample from the big data with the selection probability proportional to the importance weights.

To correct for selection bias using the proposed inverse sampling approach, we need external information about the target population, either from a census or from a probability sample, for some auxiliary variable \mathbf{x} . To formally present the idea, let (\mathbf{x}_i, y_i) be available in the big data sample (B) and $f(\mathbf{x})$ be the density for the marginal distribution of \mathbf{x} that is obtained from an external source. We assume that the auxiliary variable \mathbf{x} has a finite second moment. We are interested in estimating $\theta = E(Y)$ from the big data sample B . The first-order inclusion probability for the big data sample B is unknown.

Using the idea of importance sampling (Goffinet and Wallach, 1996; Henmi *et al.* 2007), it can be shown that

$$\hat{\theta}_{B1} = \frac{\sum_{i \in B} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i | \delta_i = 1)} \frac{f(y_i | \mathbf{x}_i)}{f(y_i | \mathbf{x}_i, \delta_i = 1)} y_i}{\sum_{i \in B} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i | \delta_i = 1)} \frac{f(y_i | \mathbf{x}_i)}{f(y_i | \mathbf{x}_i, \delta_i = 1)}}, \quad (3)$$

is asymptotically unbiased for $\theta = E(Y)$ by assuming that $f(\delta_i = 1 | \mathbf{x}_i) > 0$ for $i \in U$ almost surely. If the sampling mechanism for B is ignorable after controlling on \mathbf{x} , i.e. $P(\delta_i = 1 | \mathbf{x}_i, y_i) = P(\delta_i = 1 | \mathbf{x}_i)$, then (3) reduces to

$$\hat{\theta}_{B1} = \frac{\sum_{i \in B} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i | \delta_i = 1)} y_i}{\sum_{i \in B} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i | \delta_i = 1)}} := \sum_{i \in B} w_{i1} y_i. \quad (4)$$

The weight w_{i1} can be called importance weight, following the idea of importance sampling. If \mathbf{x}_i is a vector of stratum indicator variables, then $f(\mathbf{x}_i)/f(\mathbf{x}_i | \delta_i = 1)$ equals to $(N_h/N)/(n_h/n)$ for i in stratum h , which leads to unbiased estimation under stratified sampling.

If only $\bar{\mathbf{X}}_N = N^{-1} \sum_{i=1}^N \mathbf{x}_i$ is available, we can approximate $f(\mathbf{x})$ by $f_0(\mathbf{x})$, which minimizes the Kullback-Leibler distance

$$\min_{f_0 \in P_0} \int f_0(\mathbf{x}) \ln \left\{ \frac{f_0(\mathbf{x})}{f(\mathbf{x} | \delta = 1)} \right\} d\mathbf{x}, \quad (5)$$

where $P_0 = \{f(\mathbf{x}); \int \mathbf{x} f(\mathbf{x}) d\mathbf{x} = \bar{\mathbf{X}}_N\}$. The solution to (5) is

$$f_0(\mathbf{x}) = f(\mathbf{x} | \delta = 1) \frac{\exp(\mathbf{x}^T \boldsymbol{\lambda})}{E\{\exp(\mathbf{X}^T \boldsymbol{\lambda}) | \delta = 1\}}, \quad (6)$$

where $\boldsymbol{\lambda}$ satisfies $\int \mathbf{x} f_0(\mathbf{x}) d\mathbf{x} = \bar{\mathbf{X}}_N$, and D^T is the transpose of D . Thus, the selection probability for the second-phase selection is proportional to $\exp(\mathbf{x}^T \boldsymbol{\lambda})$, which is very close

to the exponential tilting calibration discussed in Kim (2010). Using (6), the weighted estimator in (4) reduces to

$$\hat{\theta}_{B1} = \frac{\sum_{i \in B} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\lambda}}) y_i}{\sum_{i \in B} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\lambda}})}, \quad (7)$$

where $\hat{\boldsymbol{\lambda}}$ satisfies

$$\frac{\sum_{i \in B} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\lambda}}) \mathbf{x}_i}{\sum_{i \in B} \exp(\mathbf{x}_i^T \hat{\boldsymbol{\lambda}})} = \bar{\mathbf{X}}_N. \quad (8)$$

Here, equation (8) can be called calibration equation (Wu and Sitter, 2001). Unlike the usual calibration estimation, we may ignore the sampling variability in estimating $\boldsymbol{\lambda}$ since N_B is large. When the sample size of B is large, the computation for calibration equation (8) may be challenging. In this case, one-step approximation (Kim, 2010) can be used.

Based on (7), we discuss how to select the second phase sample (B_2) of size n from the big data sample B such that $\hat{\theta}_{B2} = n^{-1} \sum_{i \in B_2} y_i$ is approximately design unbiased for $\hat{\theta}_{B1}$ in (4). The basic idea is to choose the conditional first-order inclusion probability $\pi_{i2|1} = P(i \in B_2 \mid i \in B)$ such that

$$\pi_{i2|1} = n w_{i1}, \quad i \in B, \quad (9)$$

where w_{i1} is the importance weight in (4). To guarantee

$$\pi_{i2|1} \in (0, 1], \quad i \in B, \quad (10)$$

we should choose $n \leq 1/\max_{i \in B}\{w_{i1}\}$. Once $\{\pi_{i2|1} : i \in B\}$ satisfying (9) and (10) are found, we can apply any unequal probability sampling techniques to obtain the second-phase sample; see Tillé (2006) for details on algorithms for unequal probability sampling designs.

Once the second-phase sample B_2 is obtained, we can use the sample mean of y_i in B_2 to estimate θ . The variance estimator of $\hat{\theta}_{B2}$ can be decomposed as

$$\text{Var}(\hat{\theta}_{B2}) = \text{Var}(\hat{\theta}_{B1}) + \text{Var}(\hat{\theta}_{B2} - \hat{\theta}_{B1}),$$

where the first term is of order $O(N_B^{-1})$, and the second term is of order $O(n^{-1})$. If $n/N_B = o(1)$, the first term can be safely ignored, and we only need to estimate the

second term. Since we can express

$$\hat{\theta}_{B2} = \sum_{i \in B_2} \frac{1}{\pi_{i2|1}} (w_{i1} y_i),$$

we can apply the standard variance estimation formula for the Horvitz–Thompson estimator (Horvitz and Thompson, 1952) by treating the big data as the finite population. That is, we can use

$$\hat{V} = \sum_{i \in B_2} \sum_{j \in B_2} \frac{\pi_{ij2|1} - \pi_{i2|1} \pi_{j2|1}}{\pi_{ij2|1}} \frac{w_{i1} y_i}{\pi_{i2|1}} \frac{w_{j1} y_j}{\pi_{j2|1}}$$

as a variance estimator for $\hat{\theta}_{B2}$, where $\pi_{ij2|1}$ is the joint inclusion probability for the second-phase sampling.

4 Data integration

Survey data integration is an emerging area of research, which aims to combine information from two independent surveys from the same target population. Kim *et al.* (2016) propose a new method of survey data integration using fractional imputation of Kim (2011) under the instrumental variable assumption, and Park *et al.* (2017) use a measurement error model to combine information from two independent surveys.

Survey data integration idea can be used to combine big data with survey data. Here, we assume that we have two data sources, one is a survey data (denoted by A) and the other is a big data (denoted by B) which is subject to selection bias. We assume that item \mathbf{x} is available from survey data while (\mathbf{x}, y) is available from the big data, and $n/N_B = o(1)$, where n is the sample size of A . We are interested in estimating the population mean \bar{Y}_N by combining two data sources. Because of the selection bias, the sample mean \bar{Y}_B from the big data is biased. Table 1 presents the data structure for this setup.

If both samples were probability samples, then synthetic data imputation can be used to create imputed values of y_i in the sample A . Such synthetic data imputation, or mass imputation, is also considered by Legg and Fuller (2009) and Kim and Rao (2011). When

B is a non-probability sample, Rivers (2007) proposes a mass imputation approach using nearest neighbor imputation for survey integration. That is, we can use \mathbf{x} to find the nearest neighbor in the big data sample B to create an imputed value of y_i for each element in the sample A . Once the imputed values of y_i are created for all the elements in the sample A , we can compute an imputed estimator of $\theta = E(Y)$ from the sample A . Such a method can be justified if

$$f_B(y \mid \mathbf{x}) = f(y \mid \mathbf{x}), \quad (11)$$

where $f_B(y \mid \mathbf{x})$ is the conditional density of y given \mathbf{x} for the big data sample B , and $f(y \mid \mathbf{x})$ is that for the target population. This assumption, which is called transportability, can be achieved if the selection mechanism for big data is non-informative (Pfeffermann, 1993). Because the sample A is a probability sample, the imputation estimator $\hat{\theta}_{A,I} = N^{-1} \sum_{i \in A} d_i y_i^*$ is approximately unbiased under certain conditions, where y_i^* is the imputed value of unit i , and d_i is the associated sampling weight.

Instead of using mass imputation of Rivers (2007), we propose to use propensity score weighting for the big data based on auxiliary information in the sample A . To formally describe the idea, we first assume that we can observe δ_i , the big data sample inclusion indicator, from the sample A . That is, among the elements in the sample A , it is possible to obtain the membership information from the big data sample B . For example, if the big data sample B consists of people using a certain credit card, then we can obtain δ_i from A by asking whether person i uses the credit card.

We assume that the selection mechanism of the big data sample is ignorable

$$P(\delta_i = 1 \mid \mathbf{x}_i, y_i) = P(\delta_i = 1 \mid \mathbf{x}_i), \quad i \in U,$$

and it follows a parametric model

$$P(\delta_i = 1 \mid \mathbf{x}_i) = p_i(\boldsymbol{\lambda}) \in (0, 1], \quad i \in U, \quad (12)$$

where $p_i(\boldsymbol{\lambda}) = p(\mathbf{x}_i^T \boldsymbol{\lambda})$ for some known function $p(\cdot)$ with second continuous derivatives with respect to an unknown parameter $\boldsymbol{\lambda}$, and $p_i(\boldsymbol{\lambda})^{-1} = O(N)$. Since we observe (δ_i, \mathbf{x}_i)

from the sample A , we can estimate $\boldsymbol{\lambda}$ by maximizing the pseudo log-likelihood function of $\boldsymbol{\lambda}$ given by

$$l(\boldsymbol{\lambda}) = \sum_{i \in A} d_i [\delta_i \log\{p_i(\boldsymbol{\lambda})\} + (1 - \delta_i) \log\{1 - p_i(\boldsymbol{\lambda})\}].$$

Once the pseudo maximum likelihood estimator $\hat{\boldsymbol{\lambda}}$ is obtained, then we can use a propensity score weighting estimator, that is,

$$\hat{\theta}_{B,PS} = \frac{\sum_{i \in B} p_i(\hat{\boldsymbol{\lambda}})^{-1} y_i}{\sum_{i \in B} p_i(\hat{\boldsymbol{\lambda}})^{-1}} \quad (13)$$

as a weighted estimator of θ from the big data sample B .

To discuss variance estimation of $\hat{\theta}_{B,PS}$, note that $(\hat{\boldsymbol{\lambda}}, \hat{\theta}_{B,PS})'$ is a solution to the joint estimating equation, that is,

$$U(\theta, \boldsymbol{\lambda}) \equiv \sum_{i \in B} p_i(\boldsymbol{\lambda})^{-1} (y_i - \theta) = 0, \quad (14)$$

$$S(\boldsymbol{\lambda}) \equiv \sum_{i \in A} d_i \{\delta_i - p_i(\boldsymbol{\lambda})\} g_i(\boldsymbol{\lambda}) = 0, \quad (15)$$

where $g_i(\boldsymbol{\lambda}) = \partial \logit\{p_i(\boldsymbol{\lambda})\} / \partial \boldsymbol{\lambda}$. Thus, by using the sandwich formula, we can obtain a consistent variance estimator of $\hat{\theta}_{B,PS}$; see Appendix A for details.

Remark 1 *If we can build a working outcome regression model for $E(Y \mid \mathbf{x})$, say $E(Y \mid \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$, we can construct a doubly robust estimator (Kim and Haziza, 2014) given by*

$$\hat{\theta}_{B,DR} = \frac{1}{N} \left\{ \sum_{i \in B} \frac{1}{p_i(\hat{\boldsymbol{\lambda}})} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) + \sum_{i \in A} d_i \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \right\}, \quad (16)$$

where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient based on the big data sample. We assume that an intercept term is included in \mathbf{x} . Under the model assumption (11), $\hat{\boldsymbol{\beta}}$ can be obtained by ordinary least squares. To show double robustness, let $\hat{\theta}_{A,HT} = N^{-1} \sum_{i \in A} d_i y_i$ be the Horvitz–Thompson estimator of θ from the sample A . Note that

$$\hat{\theta}_{B,DR} - \hat{\theta}_{A,HT} = \frac{1}{N} \left\{ \sum_{i \in B} \frac{1}{p_i(\hat{\boldsymbol{\lambda}})} \hat{e}_i - \sum_{i \in A} d_i \hat{e}_i \right\},$$

where $\hat{e}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$. Thus, if the model (12) is correctly specified, we have

$$E_\delta(\hat{\theta}_{B,DR} - \hat{\theta}_{A,HT}) \approx \frac{1}{N} \left(\sum_{i \in U} e_i - \sum_{i \in A} d_i e_i \right), \quad (17)$$

where $e_i = y_i - \mathbf{x}_i^\top \boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}$. The right side of (17) is design-unbiased to zero, so $\hat{\theta}_{B,DR}$ is asymptotically unbiased under model (12). On the other hand, if $E(Y | \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ is correctly specified, then,

$$\begin{aligned} \frac{1}{N} E \left\{ \sum_{i \in B} \frac{1}{p_i(\hat{\boldsymbol{\lambda}})} \hat{e}_i \mid B \right\} &\approx \frac{1}{N} \sum_{i \in B} \frac{1}{p_i(\boldsymbol{\lambda}^*)} E(\hat{e}_i \mid B), \\ \frac{1}{N} E \left(\sum_{i \in A} d_i \hat{e}_i \mid B \right) &= \frac{1}{N} \sum_{i \in U} E(\hat{e}_i \mid B), \end{aligned}$$

where $\hat{e}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\lambda}^*$ is the probability limit of $\hat{\boldsymbol{\lambda}}$. Note that $E(\hat{e}_i \mid B) = 0$ under $E(Y | \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ and MAR. Thus, we have

$$E(\hat{\theta}_{B,DR} - \hat{\theta}_{A,HT}) \approx 0, \quad (18)$$

if the outcome regression model is correctly specified. Therefore, we have established double robustness of $\hat{\theta}_{B,DR}$. Variance estimation of $\hat{\theta}_{B,DR}$ is discussed in Appendix B.

5 Simulation Study

5.1 Inverse sampling

In this simulation study, we consider the proposed inverse sampling under a simple setup. A finite population is generated by

$$y_i = 5 + 3x_i + e_i, \quad i = 1, \dots, N,$$

where $x_i \sim \text{Exp}(1)$, $e_i \sim N(0, x_i^2)$, $N = 1,000,000$, $N(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2 , and $\text{Exp}(\lambda)$ is an exponential distribution with mean λ . The inclusion indicator of the big data sample is generated by $\delta_i \sim \text{Ber}(p_i)$ independently for $i = 1, \dots, N$, where $\text{logit}(p_i) = \phi(x_i - 2)$, $\text{Ber}(p)$ is a Bernoulli distribution with success probability p , and $\text{logit}(x) = \log(x) - \log(1 - x)$ for $x \in (0, 1)$. In addition, we assume that the population mean \bar{X}_N is known. We consider two cases, $\phi = -0.2$ and $\phi = -0.5$, and we are interested in making inference for the population mean \bar{Y}_N and a proportion

$P_N = N^{-1} \sum_{i=1}^N I(y_i < 6)$, where $I(x < a) = 1$ if $x < a$ for a given number a , and 0 otherwise.

We compute the following three estimators with $n = 500$ and $n = 1,000$, respectively, and recall that n is the sample size for the second-phase sampling.

- I. Naive estimator: We use simple random sampling to get a sample of size n from the big data sample B .
- II. Calibration estimator: From the sample obtained by the naive method, we use the exponential tilting method described in Section 3 to obtain a calibration estimator using \bar{X}_N information.
- III. Proposed inverse sampling estimator: First, we obtain the important weights in (7) satisfying the calibration condition (8), and then a sample of size n is selected by probability-proportional-to-size sampling.

We conduct 10,000 Monte Carlo simulations and compare the three estimators with respect to the bias and standard error of the point estimator, the relative bias of the estimated standard error and the coverage rate of a 95% confidence interval obtained from the Wald-type method. Table 2 summarizes the simulation results. The naive estimator works poorly since it does not account for the selection bias of the big data sample. Specifically, its coverage rate decreases as the sample size gets larger, conforming the big data paradox of Meng (2018). Although the calibration estimator works better than the naive one by incorporating external information, its performance is still questionable since its variance estimator is biased when $\phi = -0.5$, that is, when the mean of the big data sample, $N_B^{-1} \sum_{i \in B} x_i$, differs significantly from \bar{X}_N . For estimating \bar{Y}_N , which is a linear function of \bar{X}_N in our simulation, the biases of the calibration estimator and the proposed inverse sampling estimator are negligible compared with the standard errors, and the coverage rates of these two methods are close to 0.95 in spite of the small bias of the estimated variance of the calibration estimator. For estimating P_N , which is not a linear function of \bar{X}_N , the biases of the calibration estimator and the proposed inverse sampling estimator are approximately the same, but they are not negligible compared

with the standard error when $\phi = -0.5$. Thus, the coverage rates of the calibration estimator and the proposed inverse sampling estimator are below 0.95. Besides, variance estimator of the proposed inverse sampling estimator is unbiased for all cases, but that of the calibration estimator becomes worse when $\phi = -0.5$.

5.2 Data integration

We use a simulation setup similar to Kim and Haziza (2014) to compare the two proposed estimators shown in (13) and (16) with a naive estimator and Rivers' method. We consider the following two outcome regression models for generating the finite population.

I. Linear model. That is,

$$y_i = 1 + x_{1,i} + x_{2,i} + \epsilon_i, \quad i = 1, \dots, N, \quad (19)$$

where $x_{1,i} \sim N(1, 1)$, $x_{2,i} \sim \text{Exp}(1)$, $\epsilon_i \sim N(0, 1)$, $N = 1,000,000$, and $(x_{1,i}, x_{2,i}, \epsilon_i)$ is pair-wise independent.

II. Nonlinear model. That is,

$$y_i = 0.5(x_{1,i} - 1.5)^2 + x_{2,i} + \epsilon_i, \quad i = 1, \dots, N, \quad (20)$$

where $(x_{1,i}, x_{2,i}, \epsilon_i)$ is the same with those in the linear model.

The sampling indicator of the big data sample is generated by $\delta_i \sim \text{Ber}(p_i)$ independently for $i = 1, \dots, N$, and we consider the following two big data propensity models.

I. Linear logistic model. That is,

$$\text{logit}(p_i) = x_{2,i}, \quad i = 1, \dots, N. \quad (21)$$

II. Nonlinear logistic model. That is,

$$\text{logit}(p_i) = -0.5 + 0.5(x_{2,i} - 2)^2, \quad i = 1, \dots, N. \quad (22)$$

The average sampling rates for the big data are about 60% under both models.

We consider the following three scenarios to generate the finite population and the big data sample.

- I. Both the outcome regression model and the big data propensity model are linear. That is, the finite population is generated by (19), and the sampling indicator of the big data sample is generated by (21).
- II. The outcome regression model is linear, and a nonlinear logistic model is used for the big data propensity model. That is, we use (19) to generate the finite population, and use (22) to generate the sampling indicator of the big data sample.
- III. The outcome regression model is nonlinear, and the big data propensity model is linear. That is, we use (20) and (21) to generate the finite population and big data sample.

The parameter of interest is the population mean \bar{Y}_N . We use simple random sampling to get an independent sample A of size n , and we consider $n = 500$ and $n = 1,000$. We compare the following methods for estimating \bar{Y}_N and the corresponding 95% confidence interval.

- I. Naive estimator. We use sample mean and sample variance of the big data sample to make inference.
- II. Rivers' method. The nearest neighbor is obtained by the Euclidean norm based on $(x_{1,i}, x_{2,i})$.
- III. The proposed propensity score (PS) weighting estimator (13) using a logistic model for $p(\cdot)$, that is, $\text{logit}\{p_i(\boldsymbol{\lambda})\} = \lambda_0 + \lambda_1 x_{2,i}$.
- IV. The proposed doubly robust (DR) estimator in (16). The working outcome regression model is $E(y_i \mid x_{1,i}, x_{2,i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i}$, and the working big data propensity model is the same as that in Method III.

For each scenario, we conduct 2,000 Monte Carlo simulations to compare the data integration estimators regarding the bias and standard error of the point estimator and the coverage rate of the 95% confidence interval obtained by the Wald-type method. Table 3 summarizes the simulation results. The naive estimator is biased since it does not account for the random mechanism for the big data sample, and its coverage rate is zero for all scenarios. Rivers' method works well in terms of the bias and coverage rate in all three scenarios. For Scenario I and Scenario III, the proposed PS estimator has the smallest standard error compared with others, and its bias and coverage rate is as good as those by Rivers' method and the proposed DR estimator. However, the proposed PS estimator is sensitive to the mis-specification of the big data propensity model, and its estimates are biased in Scenario II, where a nonlinear logistic model is used for the big data propensity model. For all three scenarios, the proposed DR estimator works better than the Rivers' method in terms of the standard errors, and both methods have approximately the same bias and coverage rate.

Remark 2 *The asymptotic variance of the Rivers' method is σ_y^2/n (Rivers, 2007), and it is consistent with the simulation results shown in Table 3 for all scenarios, where σ_y^2 is the variance of y with respect to the outcome regression model, $\sigma_y^2 = 3$ for Scenario I and II and $\sigma_y^2 = 2.75$ for Scenario III. For the proposed DR estimator, if one of the working outcome regression model and the working big data propensity model is correctly specified, the variance of $\hat{\theta}_{B,DR}$ can be estimated by the sampling variance of the imputed values $\{\mathbf{x}_i^T \boldsymbol{\beta}^* : i \in A\}$, which is $V_{B,DR} = \boldsymbol{\beta}_x^{*T} \Sigma_{xx} \boldsymbol{\beta}_x^*/n$, where $\boldsymbol{\beta}_x^*$ is the coefficient of $(x_{1,i}, x_{2,i})$ in $\boldsymbol{\beta}^*$ shown in Remark 1, and Σ_{xx} is the variance of $(x_{1,i}, x_{2,i})$; see Appendix B for details. For Scenario I and Scenario II, $V_{B,DR} = 2/n$ and $V_{B,DR} \approx 1.25/n$ for Scenario III, and the results are consistent with those shown in Table 3. Thus, the proposed DR estimator is more efficient than the Rivers' method in all three scenarios.*

6 Conclusion

Adjusting for the selection bias in big data is an important practical problem. By properly incorporating the auxiliary information from an external source, we can reduce the selection bias either by inverse sampling or by propensity score weighting. Doubly robust estimation shows good performance in the simulation study, and extension to multiple robust estimation (Chen and Haziza, 2017) seems to be a promising research area. The proposed methods implicitly assume that the selection mechanism for big data is missing at random (MAR) in the sense of Rubin (1976). If MAR assumption does not hold, then we can build a Not-Missing-At-Random model for the selection mechanism and estimate the model parameters (Chang and Kott, 2008; Riddles *et al.*, 2016).

If there is error in the matching mechanism, then misclassification errors for δ can arise, and capture-recapture experiments (Chen and Kim, 2014) can be useful in this situation. Such extensions will be topics for future research.

Acknowledgment

The authors wish to thank Professors J. N. K. Rao and Shu Yang for their constructive comments. The research was partially supported by a grant from U.S. National Science Foundation.

Appendix

A. Variance estimation of $\hat{\theta}_{B,PS}$ in (13)

We rewrite (14) and (15) as

$$\begin{aligned} U(\theta, \boldsymbol{\lambda}) &= \sum_{i=1}^N \delta_i p_i(\boldsymbol{\lambda})^{-1} (y_i - \theta), \\ S(\boldsymbol{\lambda}) &= \sum_{i=1}^N I_i d_i \{ \delta_i - p_i(\boldsymbol{\lambda}) \} g_i(\boldsymbol{\lambda}), \end{aligned}$$

where I_i is the sampling indicator for sample A , $I_i = 1$ if $i \in A$ and 0 otherwise and $g_i(\boldsymbol{\lambda}) = \partial \logit\{p_i(\boldsymbol{\lambda})\}/\partial \boldsymbol{\lambda}$. Then, we have

$$\text{Var}\{U(\theta, \boldsymbol{\lambda})\} = \sum_{i=1}^N \{1 - p_i(\boldsymbol{\lambda})\} p_i(\boldsymbol{\lambda})^{-1} (y_i - \theta)^2, \quad (\text{A.1})$$

$$\begin{aligned} \text{Var}\{S(\boldsymbol{\lambda})\} &= E[\text{Var}\{S(\boldsymbol{\lambda}) \mid A\}] + \text{Var}[E\{S(\boldsymbol{\lambda}) \mid A\}] \\ &= E[\text{Var}\{S(\boldsymbol{\lambda}) \mid A\}], \end{aligned} \quad (\text{A.2})$$

$$\begin{aligned} \text{Cov}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda})\} &= E[\text{Cov}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda}) \mid A\}] + \text{Cov}[E\{U(\theta, \boldsymbol{\lambda}) \mid A\}, E\{S(\boldsymbol{\lambda}) \mid A\}] \\ &= E[\text{Cov}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda}) \mid A\}], \end{aligned} \quad (\text{A.3})$$

where (A.1) holds since $\{\delta_i : i \in U\}$ are pair-wise independent, the second equalities of (A.2) and (A.3) hold since δ_i is independent with I_i , and $\text{Cov}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda}) \mid A\} = \sum_{i=1}^N (y_i - \theta) I_i d_i \{1 - p_i(\boldsymbol{\lambda})\} g_i(\boldsymbol{\lambda})^\top$.

Therefore, we can estimate (A.1) to (A.3) by

$$\hat{V}\{U(\theta, \boldsymbol{\lambda})\} = \sum_{i \in B} \{1 - p_i(\boldsymbol{\lambda})\} p_i(\boldsymbol{\lambda})^{-2} (y_i - \theta)^2, \quad (\text{A.4})$$

$$\begin{aligned} \hat{V}\{S(\boldsymbol{\lambda})\} &= \hat{V}\{S(\boldsymbol{\lambda}) \mid A\} \\ &= \sum_{i \in A} d_i^2 p_i(\boldsymbol{\lambda}) \{1 - p_i(\boldsymbol{\lambda})\} g_i(\boldsymbol{\lambda})^\top g_i(\boldsymbol{\lambda}), \end{aligned} \quad (\text{A.5})$$

$$\begin{aligned} \hat{C}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda})\} &= \hat{C}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda}) \mid A\} \\ &= \sum_{i \in A \cap B} d_i p_i(\boldsymbol{\lambda})^{-1} (y_i - \theta) \{1 - p_i(\boldsymbol{\lambda})\} g_i(\boldsymbol{\lambda})^\top. \end{aligned} \quad (\text{A.6})$$

Denote

$$H(\theta, \boldsymbol{\lambda}) = \begin{pmatrix} \frac{\partial U(\theta, \boldsymbol{\lambda})}{\partial \theta^\top} & \frac{\partial U(\theta, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^\top} \\ 0 & \frac{\partial S(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}^\top} \end{pmatrix}$$

to be the Hessian matrix of $[U(\theta, \boldsymbol{\lambda})^\top, S(\boldsymbol{\lambda})^\top]^\top$, and

$$\hat{V}_{U,S}(\hat{\theta}, \hat{\boldsymbol{\lambda}}) = \begin{pmatrix} \hat{V}\{U(\theta, \boldsymbol{\lambda})\} & \hat{C}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda})\} \\ \hat{C}\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda})\}^\top & \hat{V}\{S(\boldsymbol{\lambda})\} \end{pmatrix}$$

to be the variance estimator of $\{U(\theta, \boldsymbol{\lambda}), S(\boldsymbol{\lambda})\}$ based on (A.4) to (A.6).

Thus, by the sandwich formula, the variance of $(\hat{\theta}_{B,PS}, \hat{\boldsymbol{\lambda}})$ can be estimated by

$$H(\hat{\theta}, \hat{\boldsymbol{\lambda}})^{-1} \hat{V}_{U,S}(\hat{\theta}, \hat{\boldsymbol{\lambda}}) \{H(\hat{\theta}, \hat{\boldsymbol{\lambda}})^{-1}\}^\top, \quad (\text{A.7})$$

where $\hat{\theta} = \hat{\theta}_{B,PS}$, and the variance estimator of $\hat{\theta}_{B,PS}$ is the (1,1)-th element of (A.7).

B. Variance estimation of the double robust estimator

Denote

$$\tilde{\theta}_{B,DR}(\hat{\lambda}) = \frac{1}{N} \left\{ \sum_{i \in B} \frac{1}{p_i(\hat{\lambda})} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) + \sum_{i \in A} d_i \mathbf{x}_i^T \boldsymbol{\beta}^* \right\},$$

where $\boldsymbol{\beta}^*$ is the probability limit of $\hat{\boldsymbol{\beta}}$. Since $\text{Var}(\hat{\boldsymbol{\beta}}) = O(N_B^{-1})$, $\tilde{\theta}_{B,DR}(\hat{\lambda})$ is asymptotically equivalent to $\hat{\theta}_{B,DR}(\hat{\lambda})$ if $n/N_B = o(1)$.

Let $\boldsymbol{\lambda}^*$ be the probability limit of $\hat{\lambda}$, and we have

$$\tilde{\theta}_{B,DR}(\hat{\lambda}) = N^{-1} \sum_{i \in B} \frac{1}{p_i(\boldsymbol{\lambda}^*)} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) + \eta_B(\boldsymbol{\lambda}^*)^T (\hat{\lambda} - \boldsymbol{\lambda}^*) + \hat{\theta}_{A,reg} + o_p(n^{-1/2}) \quad (\text{B.1})$$

by Taylor expansion, where $\eta_B(\boldsymbol{\lambda}^*) = N^{-1} \sum_{i \in B} p_i(\boldsymbol{\lambda}^*)^{-1} \{p_i(\boldsymbol{\lambda}^*) - 1\} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) \mathbf{x}_i$ and $\hat{\theta}_{A,reg} = N^{-1} \sum_{i \in A} d_i \mathbf{x}_i^T \boldsymbol{\beta}^*$.

Note that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$. Under the model assumption (11), $\boldsymbol{\beta}^*$ is also the probability limit of $\boldsymbol{\beta}_N$, where $\boldsymbol{\beta}_N$ solves $\sum_{i=1}^N (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = 0$. Thus, we have

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p(N_B^{-1/2}), \quad (\text{B.2})$$

$$\boldsymbol{\beta}_N = \boldsymbol{\beta}^* + O_p(N_B^{-1/2}), \quad (\text{B.3})$$

where the second result holds since $\boldsymbol{\beta}_N = \boldsymbol{\beta}^* + O_p(N_B^{-1/2})$ and $N_B/N = O(1)$. Next, we wish to show

$$\eta_B(\boldsymbol{\lambda}^*) = O_p(N_B^{-1/2}), \quad (\text{B.4})$$

if one of the outcome regression model and the big data propensity model is correctly specified. Suppose that the outcome regression model is correctly specified. Then, $e_i = y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*$ is independent with \mathbf{x}_i , so (B.4) holds under mild conditions on the working big data propensity model.

If the big data propensity model is correctly specified, consider

$$\eta_B(\boldsymbol{\lambda}^*) = N^{-1} \sum_{i \in B} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) - N^{-1} \sum_{i \in B} p_i(\boldsymbol{\lambda}^*)^{-1} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) = \eta_{B,1}(\boldsymbol{\lambda}^*) - \eta_{B,2}(\boldsymbol{\lambda}^*).$$

First, note that $\sum_{i \in B} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) = 0$, and we have

$$\begin{aligned}
\eta_{B,1}(\boldsymbol{\lambda}^*) &= N^{-1} \sum_{i \in B} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) = N^{-1} \sum_{i \in B} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}) + N^{-1} \sum_{i \in B} \mathbf{x}_i^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\
&\leq O_p(N_B^{-1/2}) N^{-1} \sum_{i=1}^N \|\mathbf{x}_i\|_2 \\
&= O_p(N_B^{-1/2}), \tag{B.5}
\end{aligned}$$

where the inequality holds by (B.2), and the second equality holds if \mathbf{x}_i has a finite second moment. Now, to discuss $\eta_{B,2}(\boldsymbol{\lambda}^*)$, note that $\boldsymbol{\beta}_N$ satisfies $\sum_{i=1}^N (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_N) = 0$. Thus,

$$\begin{aligned}
\eta_{B,2}(\boldsymbol{\lambda}^*) &= N^{-1} \sum_{i \in B} p_i(\boldsymbol{\lambda}^*)^{-1} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) \\
&= N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) + O_p(N_B^{-1/2}) \\
&= N^{-1} \sum_{i=1}^N (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_N) + N^{-1} \sum_{i=1}^N \mathbf{x}_i^T (\boldsymbol{\beta}_N - \boldsymbol{\beta}^*) + O_p(N_B^{-1/2}) \\
&= O_p(N_B^{-1/2}), \tag{B.6}
\end{aligned}$$

where the last equality holds by (B.3). Thus, if the big data propensity model is correctly specified, we have shown (B.4) by (B.5) and (B.6).

Similarly, we can show that the first term of (B.1) has order $O_p(N_B^{-1/2})$ if one of the outcome regression model and the big data propensity model is correctly specified. Thus, the variance of $\tilde{\theta}_{B,DR}(\hat{\boldsymbol{\lambda}})$ can be estimated by the sampling variance of $\hat{\theta}_{A,reg}$ under the assumption $n/N_B = o(1)$.

References

- Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling, *J. Surv. Stat. Methodol.*, **1**, 90–143.
- Bethlehem, J. (2016). Solving the nonresponse problem with sample matching?, *Soc. Sci. Comput. Rev.*, **34**, 59–77.

- Breidt, F.J. & Fuller, W.A. (1993). Regression weighting for multipurpose sampling, *Sankhya B*, **55**, 297–309.
- Chang, T. & Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95**, 555–571.
- Chen, S. & Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys, *Biometrika*, **104**, 439–453.
- Chen, S. & Kim, J.K. (2014). Two-phase sampling experiment for propensity score estimation in self-selected samples, *Ann. Appl. Stat.*, **8**, 1492–1515.
- Cochran, W.G. (1977). *Sampling Techniques*, 3rd edn, John Wiley & Sons, New York.
- Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*, Cambridge, New York.
- Elliott, M. & Valliant, R. (2017). Inference for non-probability samples, *Stat. Sci.*, **32**, 249–264.
- Franke, B., Plante, J.-F., Roscher, R., Lee, E.-S.A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M.M., Grosse, R., Hendricks, D. & Reid, N. (2016). Statistical inference, learning and models in big data, *Int. Stat. Rev.*, **84**, 371–389.
- Fuller, W.A. (2009). *Sampling Statistics*, John Wiley & Sons, Hoboken.
- Goffinet, B. & Wallach, D. (1996). Optimized importance sampling quantile estimation, *Biometrika*, **83**, 791–800.
- Henmi, M., Yoshida, R. & Eguchi, S. (2007). Importance sampling via the estimated sampler, *Biometrika*, **94**, 985–991.
- Hidiroglou, M. (2001). Double sampling, *Surv. Methodol.*, **27**, 143–154.
- Hinkins, S., Oh, H.L. & Scheuren, F. (1997). Inverse sampling design algorithms, *Surv. Methodol.*, **23**, 11–21.
- Horvitz, D.G. & Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.*, **47**(260): 663–685.
- Keiding, N. & Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussions), *J. Roy. Statist. Soc. Ser. A*, **179**, 1–28.

- Kim, J.K. (2010). Calibration estimation using exponential tilting in sample surveys, *Surv. Methodol.*, **36**, 145–155.
- Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika*, **98**, 119–132.
- Kim, J.K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation, *Surv. Methodol.*, **42**, 19–40.
- Kim, J.K. & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling, *Statist. Sinica*, **24**, 375–94.
- Kim, J.K., Navarro, A. & Fuller, W.A. (2006). Replicate variance estimation after multi-phase stratified sampling, *J. Amer. Statist. Assoc.*, **101**, 312–320.
- Kim, J.K. & Rao, J.N.K. (2011). Combining data from two independent surveys: a model-assisted approach, *Biometrika*, **99**, 85–100.
- Legg, J.C. & Fuller, W.A. (2009). Two-phase sampling, in D. Pfeffermann & C.R. Rao (eds), *Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications*, North Holland, pp. 55–70.
- Meng, X.L. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and 2016 US presidential election. Submitted.
- Park, S., Kim, J.K. & Stukel, D. (2017). A measurement error model for survey data integration: combining information from two surveys, *Metron*, **75**, 345–357.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data, *Int. Stat. Rev.*, **61**, 317–337.
- Rao, J.N.K. & Molina, I. (2015). *Small Area Estimation*, 2nd edn, John Wiley & Sons, Hoboken.
- Rao, J.N.K., Scott, A.J. & Benhin, E. (2003). Undoing complex survey data structures: some theory and applications of inverse sampling, *Surv. Methodol.*, **29**, 107–128.
- Rao, J.N.K. & Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data, *Biometrika*, **82**, 453–460.
- Riddles, M.K., Kim, J.K. & Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse, *J. Surv. Stat. Methodol.*, **4**, 215–245.

- Rivers, D. (2007). Sampling for web surveys, *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Rubin, D.B. (1976). Inference and missing data, *Biometrika*, **63**(3): 581–592.
- Särndal, C.E., Cassel, C.M. & Wretman, J.H. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.
- Stukel, D. & Kott, P. (1996). Jackknife variance estimation under two-phase sampling: An empirical investigation., *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Tam, S.-M. (2015). A statistical framework for analysing big data, *Surv. Statist.*, **72**, 36–51.
- Tam, S.-M. & Clarke, F. (2015). Big data, official statistics and some initiatives by the australian bureau of statistics, *Int. Stat. Rev.*, **83**, 436–448.
- Tam, S.-M. & Kim, J.K. (2018). Big data, selection bias and ethics – an official statistician’s perspective, *Stat. J. IAOS*. Accepted for publication.
- Tillé, Y. (2006). *Sampling Algorithms*, Springer-Verlag, New York.
- Tillé, Y. (2016). Unequal probability inverse sampling, *Surv. Methodol.*, **42**, 283–295.
- Wu, C. & Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data, *J. Amer. Statist. Assoc.*, **96**, 185–193.

Table 1: Data Structure

Data	Representativeness	X	Y
A	Yes	✓	
B	No	✓	✓

Table 2: Monte Carlo bias (Bias), standard error (SE), relative bias of the estimated standard error (RB.SE) and coverage rate (CR) for different estimators based on 2,000 simulation studies. “Naive” stands for the naive estimator, “Calibration” for the calibration estimator, and “Proposed” for the proposed inverse sampling estimator. “Par.” is short for the parameter that we are interested in.

Par.	ϕ	Method	$n = 500$				$n = 1000$			
			Bias	SE	RB.SE	CR	Bias	SE	RB.SE	CR
\bar{Y}_N	-0.2	Naive	-0.27	0.133	0.01	0.48	-0.27	0.096	-0.01	0.22
		Calibration	0.00	0.070	-0.01	0.95	0.00	0.050	-0.02	0.95
		Proposed	0.00	0.146	0.01	0.95	0.00	0.104	0.00	0.95
	-0.5	Naive	-0.55	0.117	0.01	0.01	-0.55	0.083	0.00	0.00
		Calibration	0.00	0.081	-0.04	0.94	0.00	0.057	-0.03	0.94
		Proposed	0.00	0.141	0.00	0.95	0.00	0.101	0.00	0.95
P_N	-0.2	Naive	0.02	0.021	-0.01	0.83	0.02	0.015	-0.01	0.70
		Calibration	0.00	0.018	0.01	0.95	0.00	0.012	0.01	0.95
		Proposed	0.00	0.021	-0.01	0.95	0.00	0.015	0.00	0.95
	-0.5	Naive	0.05	0.021	0.00	0.45	0.04	0.015	0.01	0.16
		Calibration	-0.01	0.018	0.09	0.92	-0.01	0.012	0.09	0.89
		Proposed	-0.01	0.021	0.00	0.92	-0.01	0.014	0.02	0.90

Table 3: Monte Carlo bias (Bias), standard error (SE) and coverage rate (CR) of different data integration methods based on 2,000 simulation studies for each scenario. “Naive” stands for the naive estimator, “Rivers” for the Rivers’ method, “PS” for the proposed propensity score weighting estimator and “DR” for the proposed doubly robust estimator.

Scenario	Method	$n = 500$			$n = 1000$		
		Bias	SE	CR	Bias	SE	CR
I	Naive	0.19	0.001	0.00	0.19	0.001	0.00
	Rivers	0.00	0.077	0.95	0.00	0.054	0.95
	PS	0.00	0.023	0.95	0.00	0.016	0.95
	DR	0.00	0.063	0.95	0.00	0.044	0.95
II	Naive	-0.10	0.001	0.00	-0.10	0.001	0.00
	Rivers	0.00	0.077	0.96	0.00	0.055	0.94
	PS	0.11	0.183	0.99	0.08	0.085	1.00
	DR	0.00	0.063	0.95	0.00	0.046	0.95
III	Naive	0.19	0.001	0.00	0.19	0.001	0.00
	Rivers	0.00	0.074	0.94	0.00	0.053	0.95
	PS	0.00	0.022	0.95	0.00	0.016	0.95
	DR	0.00	0.050	0.95	0.00	0.035	0.95