

Evaluating the Impact of a Non-Probability Sample-Based Estimator in a Linear Combination with an Estimator from a Probability Sample

Journal of Official Statistics

2025, Vol. 41 (2) 649–674

© The Author(s) 2025

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0282423X251331346

journals.sagepub.com/home/jof



Andrius Čiginas¹ , Danutė Krapavickaitė²  and Vilma Nekrašaitė-Liege³ 

Abstract

In this article, the estimators based on data from independent non-probability and probability samples are combined to estimate finite population parameters. Assuming that the values of the study variable are available in both samples, the integration of the non-probability and probability samples through a composite estimator of the population total is studied. The integration is done using a linear combination of the inverse probability weighted (IPW) estimator and a design-based estimator. By evaluating the variance of the former estimator, the randomness of the underlying non-probability sample is taken into account through the distribution of the estimated propensity scores. This approach is then compared with a variance estimator based on the asymptotic variance and with a bootstrap variance estimator. The proposed linear combination is not sensitive to the misspecification of the model for the propensity scores due to the incorporated estimator of the bias of the IPW estimator. The number of Lithuanian companies possessing websites is estimated in a simulation study. By combining the sample survey data and big voluntary sample data, the properties of the introduced estimators are demonstrated numerically.

Keywords

propensity score, variance estimation, Poisson pseudo-sampling design, super-population, composite estimator

¹Vilnius University, Vilnius, Lithuania

²Lithuanian Statistical Society, Vilnius, Lithuania

³Vilnius Gediminas Technical University, Vilnius, Lithuania

Corresponding author:

Andrius Čiginas, Vilnius University, Akademijos str. 4, Vilnius LT-08412, Lithuania.

Email: andrius.ciginas@mif.vu.lt



Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

1. Introduction

Probability sampling methods are a generally accepted approach for surveying finite populations. Even for a relatively small probability sample from a finite population, valid inferences can be drawn using a known sampling design and models. Over the years, probabilistic sample-based methods have been well-developed for solving various finite population problems (Cochran 1977; Särndal et al. 1992; Tillé 2020; Wu and Thompson 2020).

However, with time passing, the situation is changing. Nowadays, probability surveys have significant drawbacks: their costs are relatively high, participation rates are decreasing, and the accuracy of estimates becomes lower. At the same time, the number of non-probability data sets is increasing. They consist of administrative data sources that record business events, data derived by sensors and machines when measuring events and situations in the physical world, and so on (Japiec et al. 2015). Compared to a probability sample, the large non-probability data sources have a lower data acquisition cost per unit. However, they also require substantial reorganization to align with a structure conducive to statistical calculations, and the (usually) unknown basis of their formation makes unbiased estimation from such samples very hard.

Nevertheless, non-probability data sets have received much attention from researchers and practitioners during the past decades, and active research is underway to utilize them effectively by combining them with probability survey data. Lohr and Raghunathan (2017) review statistical methods that have been proposed for combining information from multiple probability samples and other sources to answer research and societal questions. Zhang (2019) examines the conditions under which descriptive inference can be based directly on the observed distribution in a non-probability sample. An overview of the methods devoted to the non-probability samples is presented in Wu (2022), with a subsequent discussion in the same issue of the *Survey Methodology* journal. Lothian et al. (2019) discuss challenges arising for Central Statistical Agencies when linking disparate data sets across time, space, and sources.

The explosion of interest in non-probability samples in the twenty-first century is demonstrated graphically in Salvatore (2023). As mentioned above, one reason for this is increased data availability and lower relative costs compared to collecting data from probability samples. However, the risk of bias is obvious because of an unknown data generation mechanism. And if the selection bias is not taken into account, then even a huge non-probability sample may lead to lower accuracy results than a small probability sample (Meng 2018).

The options for correcting the non-probability sample selection bias depend on the additional information available. One of the scenarios considered in the literature is to assume that the study variable is observed only in the non-probability sample, and the independent reference probability sample contains both sampling design information on the same target population and the auxiliary variables which are common for both samples (Chen et al. 2020; Kim and Wang 2019; Yang et al. 2020). The reference probability sample is unnecessary, and analysis is technically

simpler when the values of the auxiliary variables are available for the whole population (Burakauskaitė and Čiginas 2023). This happens when access to high-quality population registers is available. Another situation encountered in practice is when the study variable is observed in both samples (Kim and Tam 2021; Rueda et al. 2023; Tam and Kim 2018). More data integration scenarios are reviewed in Yang and Kim (2020) and Rao (2021).

This paper looks at the case of non-probability and probability samples that are available independently from a common target population. Both samples contain the study variable, and the auxiliary variables are known for all elements in the frame population, which we assume is identical to the target population. An inverse probability weighting (IPW) estimator, where the estimated propensity scores are used to weigh the non-probability sample observations, is applied to estimate the population total of the study variable within this framework. A propensity score adjustment is one of the approaches used to correct the sample selection bias, where the propensity scores model the participation of the population elements in the non-probability sample (Liu et al. 2023; Wu and Thompson 2020). An alternative to this estimator is the traditional design-based estimator, using only the probability sample information and auxiliary variables.

We consider data integration through a linear combination of these two estimators, called the composite estimator. Similar composite estimators for non-probability and probability samples are applied in Elliott and Haviland (2007), Tam and Kim (2018), Zhang (2019), and Rueda et al. (2023). To optimize the combination, the variance and the possible bias of the IPW estimator should be properly evaluated, while the variance of the (approximately) unbiased design-based estimator is estimated using conventional methods. Our study considers only the bias and the variance of the estimators; we do not consider non-response and other non-sampling errors.

Variance estimators for the IPW estimator have been proposed before. Assuming a logistic regression model for the propensity scores, Chen et al. (2020) propose a simple plug-in variance estimator obtained from its asymptotic variance, and for a general parametric model for the propensity scores, Kim and Wang (2019) applied a sandwich formula to obtain a consistent variance estimator. None of these variance estimators consider the randomness effect of the unknown non-probability sample collection mechanism.

We propose the variance estimator accounting for this type of randomness through the distribution of the estimated propensity scores. Elliott and Valliant (2017) and Liu et al. (2023) give some recommendations on the resampling of non-probability samples. For comparison, we also apply the bootstrap procedure proposed by Chauvet (2007) and its algorithm presented in Mashreghi et al. (2016) to estimate the variance of the IPW estimator that takes into account the randomness of the non-probability sample. The estimator of the total and its variance estimators based on a non-probability sample are considered in Section 2.

The IPW estimator is sensitive to the misspecification of the model for the propensity scores, which may result in biased estimates. We propose to estimate the bias of the IPW estimator utilizing the data from both non-probability and

probability samples in Section 3. In Section 3, we examine the properties of the proposed estimation approach and investigate the appropriate weight of the IPW component when we take the non-probability sample randomness and the proposed bias correction into account. We provide a numerical example through a simulation experiment studying Lithuanian companies possessing websites in Section 4. The main novelty of our work is in the two proposed variance estimation methods and in the composite estimator of the population total, which demonstrates low sensitivity to the misspecification of the model for the propensity scores. The discussion and conclusions are presented in Section 5.

2. Non-Probability Samples

Let the finite population $\mathcal{U} = \{1, \dots, N\}$ consisting of N labeled elements be available. A study variable y with the fixed values y_1, \dots, y_N is defined for each population element. The parameter of interest is the total

$$t_y = \sum_{k=1}^N y_k. \quad (1)$$

Let $x^{(0)}, \dots, x^{(m)}$ be $m+1$ auxiliary variables with the values known for the whole population \mathcal{U} . For the element $k \in \mathcal{U}$, these variables attain a vector value $\mathbf{x}_k = (x_{k0}, \dots, x_{km})'$ with $x_{k0} = 1$.

A non-probability sample $B \subset \mathcal{U}$ is available. The mechanism of its formation is unknown, but the values $y_k, k \in B$, are observed and may be used for estimation of the total t_y . We assume that this sample may be considered random.

In this section, we construct the estimator of the total t_y based on the non-probability sample under certain assumptions on its origin and present some of its properties. Treating the non-probability sample as random, we propose two variance estimators for this estimator. For comparison, the variance estimator by Chen et al. (2020) will be used.

We consider the Hájek-type estimator of the total t_y :

$$\hat{t}_B = \frac{N}{\hat{N}} \sum_{k \in B} \hat{w}_k^* y_k, \text{ where } \hat{N} = \sum_{k \in B} \hat{w}_k^*, \quad (2)$$

which is based on the estimated pseudo-weights \hat{w}_k^* (a term introduced by Elliott (2009)) constructed for the non-probability sample elements.

2.1. Estimation of Pseudo-Weights

Let the number r_k be the value of the pseudo-inclusion variable r describing the belonging of the unit $k \in \mathcal{U}$ to the sample B with the value $r_k = 1$ if $k \in B$ and $r_k = 0$ otherwise. The model for the variable r will be constructed using data $\mathbf{x}_k, k \in \mathcal{U}$. Valliant (2009) notes that “a super-population model is a way of formalizing a relationship between a target variable and auxiliary data.” In our case, the

role of the target variable is played by the variable r , and we aim to establish its relationship with the auxiliary variables $x^{(0)}, x^{(1)}, \dots, x^{(m)}$. We define N random binary variables R_1, \dots, R_N with observed values r_1, \dots, r_N in the finite population \mathcal{U} . Each random variable R_k is associated with the value of the study variable y_k and the vector of auxiliary variable values \mathbf{x}_k . It is called a pseudo-inclusion indicator.

Random variables R_k are non-identically distributed according to the *Bernoulli*(π_k^*) distribution. The probability $\pi_k^* = P(R_k = 1 \mid \mathbf{x}_k, y_k)$ is called a propensity score. It is used to describe the probability that the population element $k \in \mathcal{U}$ belongs to the non-probability sample B . We are going to apply a combination of the quasi-randomization approach (Elliott and Valliant 2017), using pseudo-inclusion probabilities π_k^* , $k \in \mathcal{U}$, together with the super-population framework (Hartley and Sielken 1975).

Construction of the non-probability sample-based estimator of the total Equation (1) needs some assumptions for the pseudo-inclusion indicators R_k , similar to those in Wu (2022):

- (A1) the pseudo-inclusion indicator R_k and the variable value y_k are independent given the covariates \mathbf{x}_k , $P(R_k = 1 \mid \mathbf{x}_k, y_k) = P(R_k = 1 \mid \mathbf{x}_k)$, $k \in \mathcal{U}$;
- (A2) all probabilities π_k^* are positive: $\pi_k^* > 0$, $k \in \mathcal{U}$;
- (A3) the finite population elements enter the non-probability sample with probabilities π_k^* independently of each other, that is, the random variables R_k and R_l , $k, l \in \mathcal{U}$, $k \neq l$, are conditionally independent given \mathbf{x}_k and \mathbf{x}_l : $P(R_k = 1, R_l = 1 \mid \mathbf{x}_k, \mathbf{x}_l) = P(R_k = 1 \mid \mathbf{x}_k)P(R_l = 1 \mid \mathbf{x}_l)$.

Remark 1. Assumption (A2) ensures that all population elements can get to the non-probability sample, while assumption (A1) means a non-informative collection of the non-probability sample. Assumption (A3) implies that the non-probability sample size is random. Therefore, we consider the entry of the population elements into the sample B approximated by a Poisson sampling design if (A2) and (A3) hold (Särndal et al. 1992, p. 85). We call it a Poisson pseudo-sampling design in the context of non-probability samples.

Remark 2. Assumptions (A1), (A2), and (A3) place significant restrictions on non-probability samples, which have prompted discussion among survey statisticians. For instance, a sample of volunteers may not be missing at random and thus fail to meet the assumption (A1). A shift away from strict reliance on assumption (A1) is emerging in some recent studies (Beaumont 2020; Kim and Morikawa 2023; Liu et al. 2024), although this shift is still in its early stages. Wu (2022) dedicated Section 7 of his article to revising the assumptions. In cases where (A2) is violated, he suggests considering the “stochastic under-coverage” scenario. Assumption (A3) covers only a subset of all possible non-probability samples. For example, non-probability samples based on incomplete administrative data may satisfy (A3), whereas network and quota samples often may not satisfy it.

To estimate the propensity scores under assumptions (A1) to (A3), a super-population parametric logistic regression model for distribution of R_k is applied:

$$\pi_k^* = \pi(\mathbf{x}_k, \boldsymbol{\beta}) = \frac{\exp\{\mathbf{x}_k' \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_k' \boldsymbol{\beta}\}}, \quad k \in U, \quad (3)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_m)'$ is the vector of the real-valued parameters to be estimated. This model is fitted using the observed data $\{(r_k, \mathbf{x}_k), k \in U\}$ specifying the non-probability sample B . After the estimate $\hat{\boldsymbol{\beta}}_N = (\hat{\beta}_0, \dots, \hat{\beta}_m)'$ of the model parameter $\boldsymbol{\beta}$ is obtained, it is plugged in Equation (3) to get the estimated propensity scores

$$\hat{\pi}_k^* = \pi(\mathbf{x}_k, \hat{\boldsymbol{\beta}}_N), \quad k \in U. \quad (4)$$

Then the estimator Equation (2) with the pseudo-weights $\hat{w}_k^* = 1/\hat{\pi}_k^*$ is the IPW estimator of the total t_y .

To estimate the vector parameter $\boldsymbol{\beta}$ of the propensity scores π_k^* , the maximum likelihood (ML) method is used. Due to the conditional independence of the indicators R_k , $k \in U$, the ML method starts with defining the likelihood function

$$L(\boldsymbol{\beta}) = \prod_{k=1}^N \pi(\mathbf{x}_k, \boldsymbol{\beta})^{R_k} (1 - \pi(\mathbf{x}_k, \boldsymbol{\beta}))^{1-R_k}$$

and its log-likelihood function

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \sum_{k=1}^N (R_k \mathbf{x}_k' \boldsymbol{\beta} - \log\{1 + \exp(\mathbf{x}_k' \boldsymbol{\beta})\}).$$

The function $l(\boldsymbol{\beta})$ is continuous and has partial derivatives with respect to the components of $\boldsymbol{\beta}$ of any order. The ML estimator $\hat{\boldsymbol{\beta}}_N$ of the parameter $\boldsymbol{\beta}$ is found by solving the system of nonlinear equations

$$\frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum_{k=1}^N (R_k - \pi(\mathbf{x}_k, \boldsymbol{\beta})) x_{kj} = 0, \quad j = 0, \dots, m. \quad (5)$$

Since the estimator $\hat{\boldsymbol{\beta}}_N$ depends on the random variables R_k , $k = 1, \dots, N$, it is itself a random variable. Replacing the random variables R_k by their realizations r_k , iterative methods are applied to solve approximately the equation system (5). The Newton–Raphson algorithm is frequently used for this purpose. Alternative ways to evaluate $\boldsymbol{\beta}$ are based on estimating equations or calibration equations (Wu and Thompson 2020); however, the ML method is quite often used to fit the logistic regression model.

2.2. Variance Estimator Under the Randomized Propensity Scores

The variance of the estimator \hat{t}_B in (2) is studied in Chen et al. (2020) under the model Equation (2) for the propensity scores, where the ML estimator $\hat{\boldsymbol{\beta}}_N$ for the model parameter $\boldsymbol{\beta}$ is assumed to be consistent and is treated as fixed when

deriving the variance. We present asymptotic properties of $\hat{\boldsymbol{\beta}}_N$. Utilizing these properties, the variance estimator for the IPW estimator \hat{t}_B of the population total t_y , taking into account Poisson pseudo-sampling design and randomness in $\hat{\boldsymbol{\beta}}_N$, is constructed.

The collection of the pseudo-inclusion indicators R_1, \dots, R_N defines in our study a super-population.

Fahrmeir and Kaufmann (1985) presented the conditions that ensure the consistency and asymptotic normality of the ML estimators for the generalized linear models. Their result belongs to the field of classical statistics because they deal with a sample of independent observations.

In the paper, we use one of the generalized linear models—logistic regression model—for the super-population with the pseudo-inclusion indicators R_1, \dots, R_N . In order to get the asymptotic properties of the logistic regression model parameter estimator, the result for the generalized linear model should be adapted. We herein use the adaptation performed in Fahrmeir and Tutz (2001), Section 2.2. To state the proposition, an additional technical assumption is required:

- (A4) the components of the auxiliary vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$ are uniformly bounded, and the matrix $\mathbf{X} = (\mathbf{x}'_1, \dots, \mathbf{x}'_N)$ is of a full rank.

Proposition. If for the pseudo-inclusion indicators R_1, \dots, R_N assumptions (A1) to (A4) are valid, then the random ML estimator $\hat{\boldsymbol{\beta}}_N$ obtained in the super-population for the parameter $\boldsymbol{\beta}$ in Equation (3) has the following properties:

- (i) The probability that $\hat{\boldsymbol{\beta}}_N$ exists and is unique tends to 1 for $N \rightarrow \infty$.
- (ii) If $\boldsymbol{\beta}$ denotes the “true” value of the logistic regression model parameter, then for $N \rightarrow \infty$ we have $\hat{\boldsymbol{\beta}}_N \rightarrow \boldsymbol{\beta}$ in probability.
- (iii) The random variable $\sqrt{N}(\hat{\boldsymbol{\beta}}_N - \boldsymbol{\beta})$ converges in distribution to the random variable $Z \sim N_{m+1}(\mathbf{0}, \mathbf{i}^{-1}(\boldsymbol{\beta}))$, where $\mathbf{i}(\boldsymbol{\beta}) = \lim_{N \rightarrow \infty} N^{-1} \mathbf{I}(\boldsymbol{\beta})$ and $\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{V}(\boldsymbol{\beta})\mathbf{X}$ with the matrices

$$\mathbf{X} = \begin{pmatrix} x_{10} & \dots & x_{1m} \\ \dots & \dots & \dots \\ x_{N0} & \dots & x_{Nm} \end{pmatrix} \text{ and } \mathbf{V}(\boldsymbol{\beta}) = \begin{pmatrix} \pi_1^*(1 - \pi_1^*) & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & \pi_N^*(1 - \pi_N^*) \end{pmatrix}.$$

It follows from the proposition that for a large N , the vector parameter $\boldsymbol{\beta}$ can be approximated in the super-population by the values of the normally distributed random vector

$$\mathbf{b}(\hat{\boldsymbol{\beta}}_N) \sim \mathcal{N}_{m+1}(\hat{\boldsymbol{\beta}}_N, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_N)) \quad (6)$$

with the density function of the multivariate normal distribution denoted further by $\varphi(\hat{\boldsymbol{\beta}}_N)$. The realization of $\hat{\boldsymbol{\beta}}_N$ in the finite population should be followed by its inclusion in (6).

The IPW estimator \hat{t}_B is influenced by two dependent random components. Both components are caused by the non-probability sample collection process. One component is the direct effect of the pseudo-inclusion indicators R_k , $k = 1, \dots, N$, on the estimated pseudo-weights and the resulting estimate of the population total. The random pseudo-inclusion indicators R_k also lead to uncertainty in the ML estimator $\hat{\boldsymbol{\beta}}_N$ of the logistic regression model parameter $\boldsymbol{\beta}$. The second random component is the effect of the uncertainty in the ML estimator $\hat{\boldsymbol{\beta}}_N$ on the estimated pseudo-weights and resulting estimate of the finite population total. As noted in Remark 1, we consider the distribution of element pseudo-inclusion indicators as a Poisson pseudo-sampling design and denote it by q . The distribution of $\hat{\boldsymbol{\beta}}_N$, as well as the parameter $\boldsymbol{\beta}$ itself, is approximated by the multivariate normal distribution of $\mathbf{b}(\hat{\boldsymbol{\beta}}_N)$ given in Equation (6) and is denoted by ζ .

By replacing the value $\boldsymbol{\beta}$ in the logistic regression model Equation (3) with the vector $\mathbf{b}(\hat{\boldsymbol{\beta}}_N)$, we get the value $\tilde{\pi}_k^* = \pi(\mathbf{x}_k, \mathbf{b}(\hat{\boldsymbol{\beta}}_N))$, which approximates the propensity score π_k^* . We aim to estimate an anticipated variance $Var(\hat{t}_B)$ (Isaki and Fuller 1982), taking into account pseudo-sampling design and uncertainty in $\boldsymbol{\beta}$ estimation.

If the probabilities $\tilde{\pi}_k^*$ truthfully reflect the element belonging to the non-probability sample mechanism, then for a fixed value of $\mathbf{b}(\hat{\boldsymbol{\beta}}_N)$ —with the corresponding probabilities denoted by $\tilde{\pi}_{k|\zeta}^*$ (here we adopt the notation of Särndal et al. (1992, Chapter 8.3))—the ratio estimator \hat{t}_B in Equation (2) based on the Poisson pseudo-sampling design is approximately unbiased, and its approximate conditional variance is

$$AVar_q(\hat{t}_B | \zeta) = \sum_{k=1}^N \left(\frac{1}{\tilde{\pi}_{k|\zeta}^*} - 1 \right) (y_k - t_y/N)^2. \quad (7)$$

We estimate this variance by a modified formula appropriate for a Poisson sample (Z Liu and Valliant 2023):

$$\widehat{Var}_q(\hat{t}_B | \zeta) = \sum_{k \in B} \left(\frac{1}{\tilde{\pi}_{k|\zeta}^*} - 1 \right) \frac{(y_k - \hat{t}_B/N)^2}{\tilde{\pi}_{k|\zeta}^*}. \quad (8)$$

Taking into account two distributions, q and ζ , the anticipated variance $Var_{\zeta q}(\hat{t}_B)$ of the estimator \hat{t}_B can be decomposed by the law of the total variance (Särndal et al. 1992, p. 136):

$$Var(\hat{t}_B) = Var_{\zeta q}(\hat{t}_B) = Var_{\zeta} E_q(\hat{t}_B | \zeta) + E_{\zeta} Var_q(\hat{t}_B | \zeta). \quad (9)$$

The subscript q indicates that the variance and expectation are calculated according to the distribution of the indicators $\{R_k, k \in U\}$, and the subscript ζ means the variance and expectation by the distribution $\mathcal{N}_{m+1}(\hat{\boldsymbol{\beta}}_N, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_N))$. Due to the approximate unbiasedness of the estimator \hat{t}_B for a fixed value of the

random variable $\mathbf{b}(\hat{\boldsymbol{\beta}}_N)$, the first term in the relationship Equation (9) can be considered negligible. Then from Equation (9) we get the approximation

$$\text{Var}_{\zeta q}(\hat{t}_B) \cong E_{\zeta} \text{Var}_q(\hat{t}_B | \zeta) \cong E_{\zeta} A \text{Var}_q(\hat{t}_B | \zeta). \quad (10)$$

The expectation $\mathcal{I} = E_{\zeta} \text{Var}_q(\hat{t}_B | \zeta)$ is by the multivariate normal distribution with the density $\varphi(\hat{\boldsymbol{\beta}}_N)$ and can be written as an integral and further approximated according to (10):

$$\mathcal{I} = E_{\zeta} \text{Var}_q(\hat{t}_B | \zeta) = \int_{\mathbb{R}_{m+1}} \text{Var}_q(\hat{t}_B | \zeta) \varphi(\hat{\boldsymbol{\beta}}_N) d\hat{\boldsymbol{\beta}}_N \cong \int_{\mathbb{R}_{m+1}} A \text{Var}_q(\hat{t}_B | \zeta) \varphi(\hat{\boldsymbol{\beta}}_N) d\hat{\boldsymbol{\beta}}_N. \quad (11)$$

The Monte–Carlo approximation (Fahrmeir and Tutz 2001, Appendix A.5) of the integral gives us

$$\widehat{\text{Var}}_{\zeta q}(\hat{t}_B) = \widehat{\mathcal{I}} = \frac{1}{J} \sum_{j=1}^J \widehat{\text{Var}}_q(\hat{t}_B | \mathbf{b}^{(j)}(\hat{\boldsymbol{\beta}}_N)) \quad (12)$$

for a large integer number of repetitions J . Here the term $\widehat{\text{Var}}_q(\hat{t}_B | \mathbf{b}^{(j)}(\hat{\boldsymbol{\beta}}_N))$ is the estimator (8) of the variance $\text{Var}_q(\hat{t}_B | \zeta)$ with the propensity scores $\tilde{\pi}_k^{*(j)}$, $k \in B$, obtained by inserting the j th simulated value $\mathbf{b}^{(j)}(\hat{\boldsymbol{\beta}}_N)$ into Equation (3). Expression Equation (12) means that the anticipated variance estimator $\widehat{\text{Var}}_{\zeta q}(\hat{t}_B)$ equals empirical average of the design based estimators of variance $\widehat{\text{Var}}_q(\hat{t}_B)$ with the estimates of the parameter $\boldsymbol{\beta}$ spread around $\hat{\boldsymbol{\beta}}_N$ according to the distribution $\mathcal{N}_{m+1}(\hat{\boldsymbol{\beta}}_N, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_N))$.

Let us summarize the simulation procedure to implement the estimator Equation (12) of variance Equation (9).

Variance estimation algorithm:

1. Compute the ML estimate $\hat{\boldsymbol{\beta}}_N$ for the parameter $\boldsymbol{\beta}$ of the logistic regression model (3) replacing R_k by r_k . Calculate the propensity score estimates $\hat{\pi}_k^*$, $k \in \mathcal{U}$, from (3).
2. Do the following calculations for each $j = 1, \dots, J$: simulate $\mathbf{b}^{(j)}(\hat{\boldsymbol{\beta}}_N)$ by Equation (6) using the estimates $\hat{\pi}_k^*$, $k \in \mathcal{U}$; calculate $\tilde{\pi}_k^{*(j)}$, $k \in B$, from Equation (3) using the simulated vector $\mathbf{b}^{(j)}(\hat{\boldsymbol{\beta}}_N)$ instead of $\boldsymbol{\beta}$; evaluate the j th constituent in Equation (12).
3. Compute the average $\widehat{\mathcal{I}}$ in Equation (12).

In what follows, we call the estimator Equation (12) a smooth variance estimator. Two more estimators of the variance $Var(\hat{t}_B)$ are presented for comparison: an estimator by Chen et al. (2020), which considers the non-probability sample described only by the pseudo-inclusion indicators, and a bootstrap variance estimator, taking into account the randomness of the non-probability sample. They all are compared in the simulation study.

2.3. Variance Estimator Based on Asymptotic Variance

The IPW estimator of the population total t_y with the ML estimator of the parameter β also has been considered in Chen et al. (2020). These authors ascertained the properties of the IPW estimator \hat{t}_B given by Equation (2). They have shown that assuming the logistic regression model for the propensity scores under conditions (A1) to (A3) and some additional regularity conditions, the IPW estimator is asymptotically unbiased and derived its asymptotic variance. Chen et al. (2020) use this approximate variance to build a plug-in estimator of $Var(\hat{t}_B) = Var_q(\hat{t}_B)$. Due to completely known auxiliary data, the variance estimator has a simpler expression (Burakauskaitė and Čiginas 2023):

$$\widehat{Var}_q^{(a)}(\hat{t}_B) = \sum_{k \in B} (1 - \hat{\pi}_k^*) \left(\frac{y_k - \hat{t}_B/N}{\hat{\pi}_k^*} - \hat{\mathbf{b}}_2' \mathbf{x}_k \right)^2, \quad (13)$$

where

$$\hat{\mathbf{b}}_2' = \left\{ \sum_{k \in B} \left(\frac{1}{\hat{\pi}_k^*} - 1 \right) (y_k - \hat{t}_B/N) \mathbf{x}_k' \right\} \left\{ \sum_{k \in U} \hat{\pi}_k^* (1 - \hat{\pi}_k^*) \mathbf{x}_k \mathbf{x}_k' \right\}^{-1},$$

given the non-probability sample B .

The variance estimator Equation (13) considers the non-probability sample described only by pseudo-inclusion indicators.

2.4. Bootstrap Variance Estimator

Elliott and Valliant (2017) propose to use resampling methods to estimate the variances of the estimators based on the non-probability samples. They suggest that for the estimators like the IPW estimator \hat{t}_B in Equation (3), the bootstrap variance estimator should incorporate the variability in estimating the propensity scores and the variability caused by the sample selection mechanism in estimating the population total.

Elliott and Valliant (2017) mention that, given the estimates of pseudo-inclusion probabilities, design-based formulas may be used for point estimates and their variances. Mashreghi et al. (2016) presented the implementation of the bootstrap algorithm by Chauvet (2007) to estimate the variance of the IPW estimator Equation (2) for Poisson sampling. We have replaced the propensity scores π_k^* by their estimates and applied this algorithm to estimate $Var(\hat{t}_B)$ because this follows from

assumption (A3). Moreover, to take into account the randomness of the non-probability sample, the bootstrap variance estimator is averaged over the distribution of the estimated propensity scores.

Bootstrap variance estimation algorithm:

1. Compute the ML estimate $\hat{\beta}_N$ for the parameter β of the logistic regression model Equation (3). Calculate the propensity score estimates $\hat{\pi}_k^*$, $k \in \mathcal{U}$, from Equation (3).
2. Simulate $b(\hat{\beta}_N)$ by Equation (6) using the estimates $\hat{\pi}_k^*$, $k \in \mathcal{U}$. Calculate $\tilde{\pi}_k^*$, $k \in B$, from Equation (3) with the simulated vector $b(\hat{\beta}_N)$ in place of β .
3. Repeat the pair $(y_k, \tilde{\pi}_k^*)$, $\lfloor 1/\tilde{\pi}_k^* \rfloor$ times for all $k \in B$ to create \mathcal{U}^f , the fixed part of the bootstrap population.
4. To complete the bootstrap population, \mathcal{U}^* , construct the set \mathcal{U}^{c*} by selecting independently elements from $\{(y_k, \tilde{\pi}_k^*), k \in B\}$ with the probability $1/\tilde{\pi}_k^* - \lfloor 1/\tilde{\pi}_k^* \rfloor$ for the k th pair. Denote the bootstrap population by $\mathcal{U}^* = \mathcal{U}^f \cup \mathcal{U}^{c*} = \{(\bar{y}_k, \bar{\pi}_k^*), k \in \mathcal{U}^*\}$, where $(\bar{y}_k, \bar{\pi}_k^*)$ is the k th pair of this population corresponding to one of those in the initial sample $\{(y_k, \tilde{\pi}_k^*), k \in B\}$. In addition, each inclusion probability $\bar{\pi}_k^*$, $k \in \mathcal{U}^*$, is scaled by the factor $n_B / \sum_{l \in \mathcal{U}^*} \bar{\pi}_l^*$, ensuring that their sum across the bootstrap population remains constant.
5. Draw the bootstrap sample B^* elements independently from \mathcal{U}^* , with the probability $\bar{\pi}_k^*$ for the k th unit in \mathcal{U}^* .
6. Compute the bootstrap estimate \hat{t}_B^* by Equation (2) using the data $\{(\bar{y}_k, \bar{\pi}_k^*), k \in B^*\}$.
7. Repeat Steps 5 and 6 a large number of times, T , to obtain $\hat{t}_B^{*(1)}, \dots, \hat{t}_B^{*(T)}$. Calculate

$$\hat{V}_T^* = \frac{1}{T-1} \sum_{t=1}^T \left(\hat{t}_B^{*(t)} - \bar{t}_B^* \right)^2, \text{ where } \bar{t}_B^* = \frac{1}{T} \sum_{t=1}^T \hat{t}_B^{*(t)}.$$

8. Repeat Steps 4 to 7 a large number of times, D , to get $\hat{V}_{1T}^*, \dots, \hat{V}_{DT}^*$. Calculate the estimate

$$\bar{V}_{DT}^* = \frac{1}{D} \sum_{d=1}^D \hat{V}_{dT}^*.$$

9. Repeat Steps 2 to 8 a large number of times, L , to get $\bar{V}_{1DT}^*, \dots, \bar{V}_{LDT}^*$. Calculate the estimate

$$\widehat{Var}^*(\widehat{t}_B) = \frac{1}{L} \sum_{l=1}^L \bar{V}_{lDT}^* \quad (14)$$

of the variance $Var(\widehat{t}_B)$ for the IPW estimator Equation (2).

The asymptotic unbiasedness of this variance estimator, based on estimated pseudo-inclusion probabilities, remains unestablished and could be a focus of future research.

3. Design-Based and Composite Estimators of the Total and Estimation of Their Variances

The mechanism by which an element is included in the non-probability sample is unknown. Therefore, even a carefully constructed IPW estimator may not sufficiently correct the selection bias of this sample. If an additional, albeit much smaller, independently selected probability sample is available in which the study variable is observed, it provides supplementary information about the population and can be used to deal with the possibly biased IPW estimator Equation (2). For this purpose, we integrate both samples through a linear combination of the IPW and design-based estimators.

Let us assume that, independently of sample B , a probability sample A is selected from the same population \mathcal{U} according to the sampling design $p(\cdot)$, and the values y_k , $k \in A$, are observed. The samples A and B may overlap. Let I_k be the selection indicator for a unit $k \in \mathcal{U}$, selected to the sample A with the value $I_k = 1$ if $k \in A$ and $I_k = 0$ otherwise.

The design-based estimator of the population total t_y is defined as:

$$\widehat{t}_A = \sum_{k \in A} w_k y_k, \quad (15)$$

which is at least approximately unbiased, according to the randomness induced by the probability sampling design $p(\cdot)$. This estimator may be the Horvitz–Thompson estimator with exactly or approximately known inclusion probabilities $\pi_k = P(I_k = 1)$ and weights $w_k = 1/\pi_k$, $k \in A$, or another estimator using auxiliary data (Särndal et al. 1992); or the Hájek estimator

$$\widehat{t}_A = \frac{N}{\widehat{N}} \sum_{k \in A} w_k y_k \text{ with } \widehat{N} = \sum_{k \in A} w_k.$$

The integration of both samples—non-probability and probability—is done through the composite estimator of the total t_y :

$$\widetilde{t}_y^c = \widetilde{t}_y^c(\alpha) = \alpha \widehat{t}_B + (1 - \alpha) \widehat{t}_A \quad (16)$$

with the coefficient $0 \leq \alpha \leq 1$ minimizing the mean squared error (MSE) of the estimator $\widetilde{t}_y^c(\alpha)$. Since the samples A and B are selected independently, the estimators

\hat{t}_A and \hat{t}_B are independent as well, implying that $Cov(\hat{t}_A, \hat{t}_B) = 0$. Assuming further that the estimator \hat{t}_A is unbiased, the MSE can be expressed as

$$MSE(\hat{t}_y^c(\alpha)) = \alpha^2 MSE(\hat{t}_B) + (1 - \alpha)^2 Var(\hat{t}_A), \quad (17)$$

where $MSE(\hat{t}_B) = Var(\hat{t}_B) + (Bias(\hat{t}_B))^2$ with a possibly non-negligible bias component

$$Bias(\hat{t}_B) = E(\hat{t}_B) - t_y. \quad (18)$$

The minimum value of the function $MSE(\hat{t}_y^c(\alpha))$ in Equation (17) is attained at $\alpha = \alpha_0$ for

$$\alpha_0 = \frac{Var(\hat{t}_A)}{Var(\hat{t}_A) + MSE(\hat{t}_B)}, \quad (19)$$

and the coefficient α_0 has to be estimated from the data available. Here, the variance $Var(\hat{t}_A)$ is readily estimated using the standard design-based methods (Särndal et al. 1992; Wolter 2007), while the estimation of the variance $Var(\hat{t}_B)$ depends on the specific choice of the pseudo-weights \hat{w}_k^* in Equation (2) and other assumptions as considered in Section 2. If the pseudo-weight \hat{w}_k^* construction method correctly reflects the population unit involvement in the non-probability sample mechanism, then an approximately unbiased estimator \hat{t}_B can be expected. Otherwise, the estimator \hat{t}_B can be significantly biased. We further propose an estimator of the bias $Bias(\hat{t}_B)$ to evaluate the optimal coefficient α_0 Equation (19) properly.

Let us study the variability of the composite estimator Equation (16) of the population total. Three different estimators of the variance $Var(\hat{t}_B)$ for the IPW estimator Equation (2) are considered: the new estimator $\widehat{Var}_{\zeta q}(\hat{t}_B)$ given by Equation (12), the estimator $\widehat{Var}_q^{(a)}(\hat{t}_B)$ of Chen et al. (2020) given by Equation (13), and the bootstrap estimator $\widehat{Var}^*(\hat{t}_B)$ from Equation (14). Denote further by $\widehat{Var}(\hat{t}_B)$ any of these estimators.

Since the study variable is observed in the probability sample A , the true population total t_y in the expression of the bias Equation (18) is at least approximately unbiasedly estimated using \hat{t}_A , while \hat{t}_B is taken as the estimator of the population parameter $E(\hat{t}_B)$. Then the estimator

$$\widehat{Bias}(\hat{t}_B) = \hat{t}_B - \hat{t}_A \quad (20)$$

of the bias is at least approximately unbiased, with the variance $Var(\widehat{Bias}(\hat{t}_B)) = Var(\hat{t}_B) + Var(\hat{t}_A)$ due to the independence of the estimators \hat{t}_A and \hat{t}_B . The bias estimator Equation (20) is suggested also by Elliott and Haviland (2007).

Then, after the estimator $\widehat{Var}_p(\widehat{t}_A)$ of the variance $Var(\widehat{t}_A) = Var_p(\widehat{t}_A)$ of the design-based estimator \widehat{t}_A in Equation (15) is chosen, the optimal coefficient α_0 in Equation (19) for the composition Equation (16) is estimated by

$$\widehat{\alpha}_0 = \frac{\widehat{Var}_p(\widehat{t}_A)}{\widehat{Var}_p(\widehat{t}_A) + \widehat{Var}(\widehat{t}_B) + (\widehat{t}_B - \widehat{t}_A)^2}.$$

Finally, the composite estimator

$$\widehat{t}_y^c = \widehat{t}_y^c(\widehat{\alpha}_0) = \widehat{\alpha}_0 \widehat{t}_B + (1 - \widehat{\alpha}_0) \widehat{t}_A \quad (21)$$

of the population total t_y is obtained, and its MSE is estimated by

$$\widehat{MSE}(\widehat{t}_y^c) = \widehat{\alpha}_0 (\widehat{Var}(\widehat{t}_B) + (\widehat{t}_B - \widehat{t}_A)^2).$$

This formula is derived by inserting $\widehat{\alpha}_0$ into (17). The more the estimator \widehat{t}_B is biased or highly fluctuating, the lower its coefficient $\widehat{\alpha}_0$ in the composition Equation (21). The efficiency of the composite estimator Equation (21) of t_y for different choices of the variance estimator $\widehat{Var}(\widehat{t}_B)$ is examined in the simulation study.

The estimator \widehat{t}_B , based on the non-probability sample, is biased but has low variance. In contrast, the estimator \widehat{t}_A , supported by the probability sample, is unbiased but exhibits relatively high variance. By combining these two estimators, the composite estimator “borrows strength” from both, similar to techniques used in small area estimation (Rao and Molina 2015). This approach mitigates the weaknesses of the individual components, leading to an estimator with comparatively lower variance and reduced bias. The effectiveness of this method is demonstrated in the simulation study.

4. Numerical Experiment

A numerical study aims to compare the estimators considered, show the alternation of their accuracy when the researcher fails to specify the propensity score model precisely enough and demonstrate the composite estimator’s usefulness regardless of the probability sample size. The essence of the numerical experiment is to simulate the problem environment by repeatedly generating the non-probability and probability samples, estimating the population total and the variances of its estimators by several methods, and studying the distributions of the estimators obtained. The results allow us to compare the accuracy of the estimators and show the conditions under which any of these estimators can be efficient.

4.1. Simulation Data

The data set used in the simulations is constructed using three data sets (sources) with the same record identifier. The first is the probability sample data set from

the Lithuanian Information and Communication Technology (ICT) survey, with the binary study variable y indicating whether the enterprise has a website. For its selection, stratified simple random sampling is used. The values of other variables, such as the number of employees and the indicators of economic activity, are known for the whole survey population of size $N = 13\,884$. One more completely known auxiliary variable y^* is derived from the second, web-scraped data set. This binary variable approximates the study variable y quite well. The third data set, containing values of the variable y , is provided by a private company. This data source is a voluntary non-probability sample covering about 56% of the survey population.

The pseudo-real population \mathcal{U} of the real size $N = 13\,884$ is constructed by linking these three data sets. The y values in the probability survey sample are supplemented with observations in the non-probability sample, and further information is gathered about the remaining unknown values. In the population obtained, the contingency coefficient between y and y^* , measuring the relationship strength between these variables, equals 0.55 (while the maximal possible value of the contingency coefficient for 2×2 frequency table is 0.7071). The auxiliary variables $x^{(0)}, \dots, x^{(m)}$ include the web-scraped variable y^* , the number of employees, and some variables indicating economic activity.

4.2. Simulation Procedure

The simulation procedure consists of the following steps:

1. The logistic regression model Equation (3) is fitted using the linked non-probability sample. Then, the estimates Equation (4) of the propensity scores are obtained for all population elements. They are used further as Poisson sample selection probabilities.
2. Two independent samples are constructed from the pseudo-real population \mathcal{U} . The first is a stratified simple random sample A mimicking an original ICT survey sampling design. For simulation purposes, two versions of the probability samples of different sizes are considered: sample A_1 of size 3 091 and a smaller sample A_2 of size 1 052. Two different approaches are employed to collect the second sample B . In the first case, sample B_1 is selected using the Poisson sampling design, with the selection probabilities obtained in Step 1, with the expected sample size equal to 7 750. In the second case, a stratified simple random sample B_2 of size 7 742 is selected after the pseudo-real population is divided into strata according to the enterprise size (5 strata by the number of employees).

Three combinations of probability and non-probability samples are considered:

- (a) Case 1. A_1 is used as sample A , and B_1 is used as sample B . This combination corresponds to an existing situation investigating whether integrating both samples improves estimation accuracy.
 - (b) Case 2. A_2 is used as sample A , and B_1 is used as sample B . This combination uses a probability sample that is three times smaller (cheaper), allowing us to investigate the usefulness of the probability sample in reducing the biases of the estimates obtained from the non-probability sample.
 - (c) Case 3. A_2 is used as sample A , and B_2 is used as sample B . This case considers a scenario where the probability sample size is small, and the non-probability sample is represented by a sample that does not satisfy assumptions (A1) and (A3) and has a structure that is not harmonized with the estimation methodology.
3. Using the generated samples A and B , the following estimates are calculated:
 - (a) A set of covariates is chosen from the collection $x^{(0)}, \dots, x^{(m)}$, and a logistic regression model (3) is applied to the non-probability sample B to get the estimates $\hat{\pi}_k^*$, $k \in \mathcal{U}$, of the propensity scores. These estimates are then used to evaluate the estimator \hat{t}_B of the population total t_y and three estimators of the variance $Var(\hat{t}_B)$:
 - the proposed smooth estimator Equation (12), calculated with $J = 300$;
 - the bootstrap estimator Equation (14) with parameters $T = 500$, $D = 10$, and $L = 20$;
 - the variance estimator Equation (13) by Chen et al. (2020).
 - (b) Using the probability sample A , the population total t_y is estimated by applying the regression estimator \hat{t}_A^{reg} based on the auxiliary variable y^* .
 - (c) The Hájek estimator of the total, calculated from sample A , serves as a benchmark in the simulations.
 - (d) The composite estimators Equation (21) of t_y are calculated using each of the estimates of variance $Var(\hat{t}_B)$ listed in (a).
 - (e) The naive estimator $N \sum_{k \in B} y_k / N_B$ of t_y is calculated, where N_B is the size of sample B .
 4. Steps 2 and 3 are repeated $R = 1000$ times. The numerical characteristics, like the average over the repetitions, are calculated for each estimator of interest.

The selection of covariates for the logistic regression model in Step 3 is needed to imitate a situation where the data are not well-known to the researcher. The following three logistic regression models of different quality levels are used:

- I The model includes all possible covariates that are the same as those used to estimate the pseudo-inclusion probabilities in Step 1.
- II The model incorporates all covariates, except for the number of employees, which has a smaller impact on the values of the study variable y than the variable y^* .
- III The model includes all the covariates without the variable y^* .

Table 1. Quality Classification of Logistic Regression Models.

Model	Akaike criterion	Model quality
Model I	17,504	Strong
Model II	17,825	Weak
Model III	18,742	Poor

The quality comparison of the logistic regression models is performed using the Akaike information criterion and presented in Table 1. The model with a lower value of this criterion fits the data better.

According to the values of the Akaike criterion, the logistic regression models are classified as strong, weak, and poor.

4.3. Simulation Results

The numerical characteristics of the simulated estimates, namely their empirical means, biases, standard deviations, and root mean square errors (*RMSE*), are obtained for both estimates of the population total and variance: where $\hat{\theta}^{(r)}$, $r = 1, \dots, R$, are the realizations of $\hat{\theta}$, while the population parameter θ denotes t_y or $Var(\hat{t}_B)$.

First of all, the estimators of the variance $Var(\hat{t}_B)$ based on the non-probability samples used in Cases 1 to 3 are graphically overviewed. The variance estimates are the same for Cases 1 and 2 because they are applied to the same non-probability sample and are presented in Figure 1. For Case 3, the variance estimates are presented in Figure 2.

The vertical lines in Figures 1 and 2 represent the empirical variance

$$Var_{emp}(\hat{t}_B) = \frac{1}{R-1} \sum_{r=1}^R \left(\hat{t}_B^{(r)} - Mean(\hat{t}_B) \right)^2, \tag{22}$$

where $\hat{t}_B^{(r)}$, $r = 1, \dots, R$, are the IPW estimates obtained by (2), and $Mean(\hat{t}_B)$ is their average. The proposed smooth variance estimator is expected to be greater than the empirical variance because it accounts for both the variability in estimating the propensity scores and the variability caused by the sample collection mechanism in estimating the population total. In contrast, the empirical variance considers the non-probability sample described only by pseudo-inclusion indicators.

Considering Figure 1, for the poor model, all variance estimators, including the proposed smooth estimator Equation (12), are higher than the empirical variance Equation (22), behave similarly and have higher variability compared to other model cases. For the strong and weak model cases, the smooth estimator behaves similarly to the bootstrap estimator, while the variance estimates by Chen et al. (2020) are, on average, lower.

In Case 3 (Figure 2), as compared to Cases 1 and 2 (Figure 1), we observe the same trends in the variance estimates, except for the fact that the bootstrap

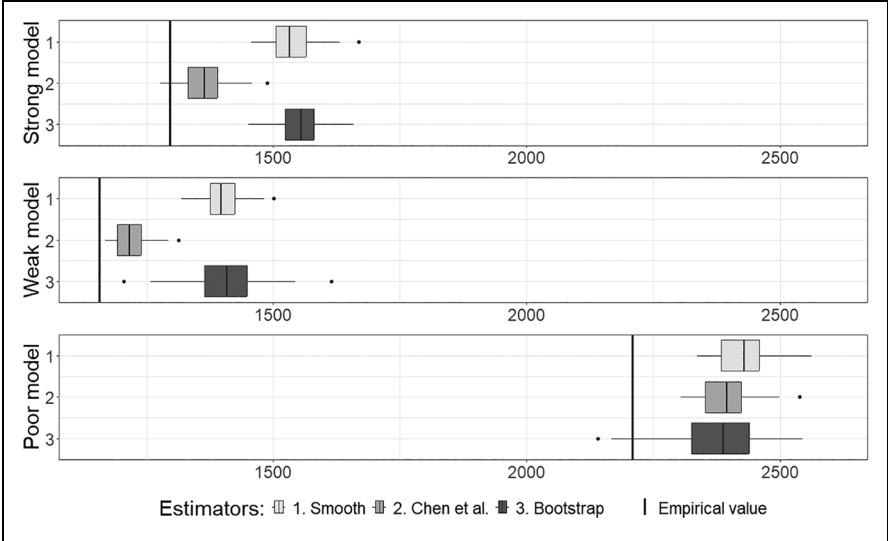


Figure 1. Variance estimates for the estimator \hat{t}_B in Cases 1 and 2.

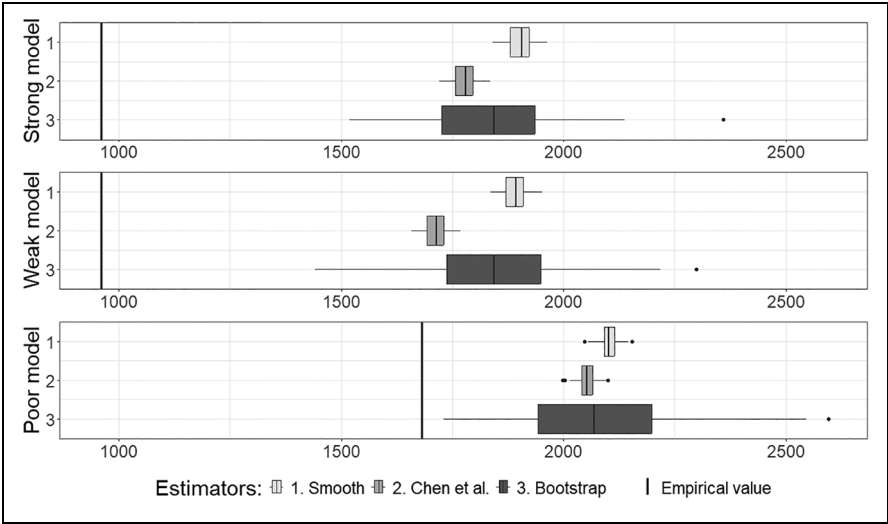


Figure 2. Variance estimates for the estimator \hat{t}_B in Case 3.

variance estimates have a much wider spread than other variance estimates. This difference may be because, in Case 3, the non-probability sample does not satisfy assumptions (A1) and (A3), and the bootstrap variance estimation is more sensitive to this factor due to resampling.

Table 2. Accuracy Measures of the Estimators of Variance for \hat{t}_B .

Model	Estimator	Cases 1 and 2		Case 3	
		Mean	SD	Mean	SD
Strong	Smooth, $\widehat{Var}_{\zeta q}(\hat{t}_B)$	1 539	41	1 903	26
	Chen et al., $\widehat{Var}_q^{(a)}(\hat{t}_B)$	1 365	40	1 779	26
	Bootstrap, $\widehat{Var}^*(\hat{t}_B)$	1 557	46	1 841	159
Weak	Smooth, $\widehat{Var}_{\zeta q}(\hat{t}_B)$	1 399	34	1 891	26
	Chen et al., $\widehat{Var}_q^{(a)}(\hat{t}_B)$	1 217	34	1 714	24
	Bootstrap, $\widehat{Var}^*(\hat{t}_B)$	1 403	70	1 838	155
Poor	Smooth, $\widehat{Var}_{\zeta q}(\hat{t}_B)$	2 424	49	2 099	19
	Chen et al., $\widehat{Var}_q^{(a)}(\hat{t}_B)$.	2 391	49	2 049	18
	Bootstrap, $\widehat{Var}^*(\hat{t}_B)$	2 385	88	2 079	185

The means and standard deviations of the variance estimates in Table 2 numerically express the situation presented in Figures 1 and 2.

The results in Table 2 confirm that in Case 3, the *SD* of the bootstrap variance estimator is comparatively large. The Chen et al. (2020) variance estimator consistently yields the lowest mean estimates across all models and cases, as it does not account for the variability caused by the sample collection mechanism. The smooth variance estimator generally produces higher mean estimates, while its *SD* remains similar to that of the Chen et al. (2020) estimator. This suggests that incorporating two sources of variability increases the value of the variance estimator without inflating its variability.

The results for all estimators of the population total t_y are presented in Figures 3 to 5, which correspond to Cases 1 to 3. Here Composite 1 denotes the composite estimator (21), with the coefficient $\hat{\alpha}_0$ using the smooth estimator of variance, Composite 2 uses the variance estimator of Chen et al. (2020), and Composite 3 is based on the bootstrap variance estimator. The estimator \hat{t}_B is denoted by NP.

Figures 3 to 5 depict the positions of the simulated estimates relative to the known true value $t_y = 10\,672$ of the population total, indicated by a straight line. The naive estimates of the total, using the non-probability sample B as a simple random sample from the population, exhibit a significant bias across all cases. The estimates \hat{t}_B (NP estimates) demonstrate a narrow spread for all cases, regardless of which logistic regression model is utilized. In Cases 1 and 2, where the role of the non-probability sample is played by a sample selected using the Poisson sampling design with the pseudo-selection probabilities obtained in Step 1, the estimator \hat{t}_B is nearly unbiased for the strong model, exhibits some bias for the weak model, and demonstrates a significant bias for the poor model. When the role of the non-probability sample is played by a sample selected in a different manner in Case 3,

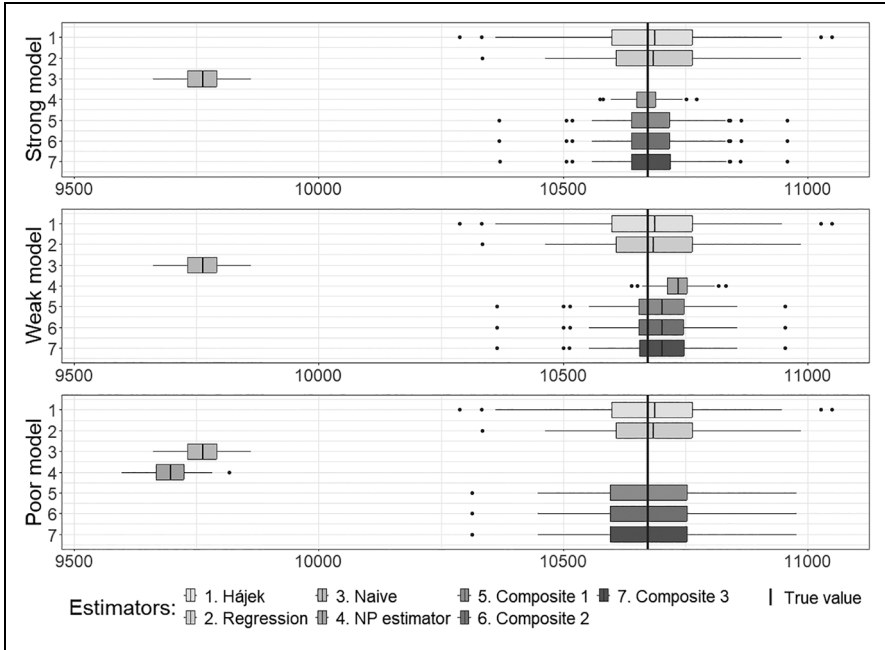


Figure 3. Estimates of the population total t_y in Case 1.

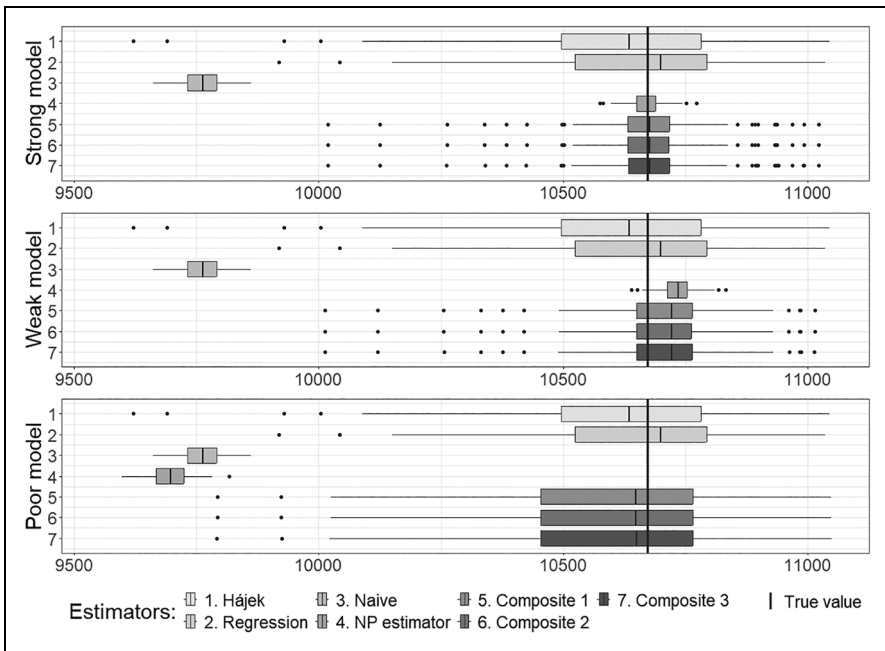


Figure 4. Estimates of the population total t_y in Case 2.

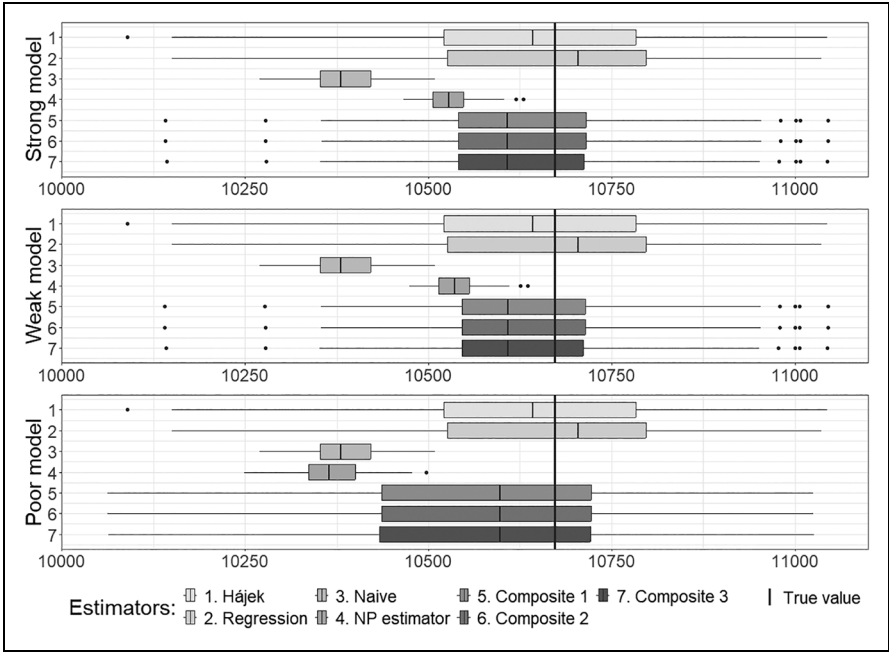


Figure 5. Estimates of the population total t_y in Case 3.

the bias of the estimator \hat{t}_B decreases (compared to the naive estimator); nevertheless, it still remains considerable. The behavior of composite estimates remains consistent regardless of the method used to estimate the optimal weight α_0 . In all cases, their spread is narrower than that of the regression estimates applied to the probability sample for the strong and weak models; however, they do not outperform regression estimates when the quality of the logistic regression model for the propensity scores is poor.

In the case where A_1 is used as sample A and B_2 is used as sample B , the same tendencies as in Figure 5 are observed. Due to the larger probability sample size and the lower variance of the unbiased estimator, the spread of the estimates is slightly narrower. The figure for this case is not shown.

The smooth variance estimator is efficient when assumptions (A1) and (A3) for the pseudo-inclusion indicators R_k hold, and the logistic regression model used to estimate pseudo-inclusion probabilities, $\hat{\pi}_k$, is sufficiently strong (well specified). As the quality of this model declines, the bias in the smooth variance estimator increases. However, in our application, a composite estimator alleviates this issue across any variance estimator, as the bias of the estimator \hat{t}_B is notably larger than its variance due to the relatively large non-probability sample.

The numerical summary of Figures 3 to 5 is provided in Tables 3 to 5, respectively.

Table 3. Accuracy Measures for the Estimators of the Population Total $t_y = 10\,672$ in Case 1.

Model	Estimator	Mean	Bias	SD	RMSE
Strong	Hájek	10 676	4	149	149
	Regression	10 684	12	111	112
	Naïve	9 761	−911	46	912
	NP estimator	10 671	−1	36	36
	Composite 1	10 682	10	82	82
	Composite 2	10 682	10	81	82
Weak	Composite 3	10 682	10	82	82
	NP estimator	10 734	62	34	71
	Composite 1	10 700	28	84	88
	Composite 2	10 700	28	84	88
Poor	Composite 3	10 700	28	84	88
	NP estimator	9 695	−977	47	978
	Composite 1	10 673	1	113	113
	Composite 2	10 673	1	113	113
	Composite 3	10 673	1	113	113

The accuracy measures in Tables 3 to 5 corroborate the observed behavior (refer to Figures 3–5, respectively) of all three composite estimators, indicating their insensitivity to the choice of the variance estimator based on the non-probability sample. For the high-quality (strong) logistic regression model, the NP estimator based solely on the non-probability sample achieves the lowest *RMSE* for all cases. However, this estimator exhibits a significant bias under the poor model. Simulation results demonstrate that composite estimators are much more stable due to the bias of the NP estimator incorporated into their expression. Even with a poor model for the propensity scores, it is possible to obtain an almost unbiased composite estimator, albeit with an *RMSE* comparable to that of the regression estimator applied to the probability sample. In the case of the weak model, composite estimators exhibit a smaller bias than the NP estimator, but their *RMSE*s are higher. It is noteworthy that the accuracy characteristics of composite estimators remain consistent regardless of the differences observed between the variance estimates for the NP estimator of the total (see Table 2).

5. Conclusions and Discussion

The estimates of the population total based on a non-probability sample usually are biased. The nature of the non-probability sample is typically unknown; however, there is no reason to presume that it is not random in some specific applications. One of the ways to decrease the bias of the estimator of the total using a non-probability sample is to assume the probabilistic nature of this sample and use it for estimation and inferences.

The assumption of an independent entry of elements into the non-probability sample with probabilities estimated by the propensity scores from the logistic

Table 4. Accuracy Measures for the Estimators of the Population Total $t_y = 10\,672$ in Case 2.

Model	Estimator	Mean	Bias	SD	RMSE
Strong	Hájek	10 700	28	338	339
	Regression	10 704	32	248	250
	Naïve	9 761	−911	46	912
	NP estimator	10 671	−1	36	36
	Composite 1	10 690	18	176	177
	Composite 2	10 690	18	176	177
Weak	Composite 3	10 691	19	177	178
	NP estimator	10 734	62	34	71
	Composite 1	10 714	42	176	181
	Composite 2	10 714	42	176	181
Poor	Composite 3	10 714	42	176	181
	NP estimator	9 695	−977	47	978
	Composite 1	10 644	−28	265	267
	Composite 2	10 644	−28	265	267
	Composite 3	10 644	−28	265	267

Table 5. Accuracy Measures for the Estimators of the Population Total $t_y = 10\,672$ in Case 3.

Model	Estimator	Mean	Bias	SD	RMSE
Strong	Hájek	10 700	28	338	339
	Regression	10 704	32	248	250
	Naïve	10 392	−280	41	283
	NP estimator	10 533	−139	31	143
	Composite 1	10 637	−35	192	195
	Composite 2	10 637	−35	192	195
Weak	Composite 3	10 636	−36	192	195
	NP estimator	10 540	−132	31	135
	Composite 1	10 639	−33	190	193
	Composite 2	10 639	−33	190	193
Poor	Composite 3	10 638	−34	190	193
	NP estimator	10 374	−298	41	301
	Composite 1	10 605	−67	225	235
	Composite 2	10 605	−67	225	235
	Composite 3	10 604	−68	225	235

regression model is well-suited for this purpose. The accuracy of the IPW estimator of the total depends on how well the propensity score model is specified. As the simulation results show, if the model adequately reflects the mechanism of formation of a non-probability sample, then the IPW estimator may have an insignificant bias. Otherwise, the IPW estimator is biased, and a probability sample with an approximately unbiased estimator of the total is needed. The proposed composite estimator with the bias component included in its coefficient gives nearly unbiased estimates in the case of any logistic regression model for the propensity scores.

However, it is not enough to assume a probabilistic sample pseudo-selection mechanism for a non-probability sample. One should also consider the possible randomness in this mechanism, as it is done for the proposed smooth estimator of variance. This estimator averages the known estimator of the variance of the ratio estimator over the distribution of propensity score estimates. According to the simulation results, if the model for the propensity scores is not poor, then this average value is close to the value of the bootstrap estimator of variance, and its spread is not wider than that for the bootstrap. The advantage of the smooth variance estimator over the bootstrap variance estimator becomes even more pronounced when the non-probability sample formation mechanism significantly differs from that described by the logistic regression model.

In our setup, the values of the study variable are available in both non-probability and probability samples. In practice, if the collection of probability sample data is expensive, then a small probability sample may be selected and utilized through composite estimation to decrease the bias of the IPW estimate. The simulation study demonstrates that the composite estimator is more efficient than the estimator based on a small probability sample and performs comparably to the estimator based on a large probability sample. Moreover, regardless of the probability sample size, composite estimators significantly improve the accuracy of other estimators of the population total if the underlying logistic regression model is well-specified.

Further research could focus on integrating probability and non-probability samples over time, addressing non-sampling errors, and estimating nonlinear population parameters while accounting for the randomness of the non-probability sample.


Acknowledgement


We sincerely thank three anonymous Referees and a Guest Editor for their valuable comments, which have brought attention to important details and significantly improved the quality of this article.


Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Andrius Čiginas  <https://orcid.org/0000-0001-8509-5034>

Danutė Krapavickaitė  <https://orcid.org/0000-0002-7638-6697>

Vilma Nekrašaitė-Liegė  <https://orcid.org/0000-0002-4922-0851>

References

- Beaumont, J.-F. 2020. "Are Probability Surveys Bound to Disappear for the Production of Official Statistics?" *Survey Methodology* 46 (1): 1–28.

- Burakauskaitė, I., and A. Čiginas. 2023. "An Approach to Integrating a Non-Probability Sample in the Population Census." *Mathematics* 11 (8): 1782–95. DOI: <https://doi.org/10.3390/math11081782>.
- Chauvet, G. 2007. "Méthodes de bootstrap en population finie." PhD thesis, Université de Rennes 2.
- Chen, Y., P. Li, and C. Wu. 2020. "Doubly Robust Inference with Nonprobability Survey Samples." *Journal of the American Statistical Association* 115 (532): 2011–21. DOI: <https://doi.org/10.1080/01621459.2019.1677241>.
- Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. John Wiley & Sons.
- Elliott, M. N., and A. Haviland. 2007. "Use of a Web-Based Convenience Sample to Supplement a Probability Sample." *Survey Methodology* 33 (2): 211–5.
- Elliott, M. R. 2009. "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights." *Survey Practice* 2 (6): 1–7. DOI: <https://doi.org/10.29115/SP-2009-0025>.
- Elliott, M. R., and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2): 249–64. DOI: <https://doi.org/10.1214/16-sts598>.
- Fahrmeir, L., and H. Kaufmann. 1985. "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models." *The Annals of Statistics* 13 (1): 342–68. DOI: <https://doi.org/10.1214/aos/1176346597>.
- Fahrmeir, L., and G. Tutz. 2001. *Multivariate Statistical Modelling Based on Generalized Linear Models*. 2nd ed. Springer. DOI: <https://doi.org/10.1007/978-1-4757-3454-6>.
- Hartley, H. O., and R. L. Sielken, Jr. 1975. "A 'Super-Population Viewpoint' for Finite Population Sampling." *Biometrics* 31 (2): 411–22. DOI: <https://doi.org/10.2307/2529429>.
- Isaki, C. T., and W. A. Fuller. 1982. "Survey Design Under the Regression Superpopulation Model." *Journal of the American Statistical Association* 37 (377): 89–96. DOI: <https://doi.org/10.2307/2287773>.
- Japac, L., F. Kreuter, M. Berg, P. Biemer, P. Decker, C. Lampe, J. Lane, C O'Neil, and A. Usher. 2015. "Big Data in Survey Research: AAPOR Task Force Report." *Public Opinion Quarterly* 79 (4): 839–80. DOI: <https://doi.org/10.1093/poq/nfv039>.
- Kim, J. K., and K. Morikawa. 2023. "An Empirical Likelihood Approach to Reduce Selection Bias in Voluntary Samples." *Calcutta Statistical Association Bulletin* 75 (1): 8–27. DOI: <https://doi.org/10.1177/00080683231186488>.
- Kim, J. K., and S.-M. Tam. 2021. "Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference." *International Statistical Review* 89 (2): 382–401. DOI: <https://doi.org/10.1111/insr.12434>.
- Kim, J. K., and Z. Wang. 2019. "Sampling Techniques for Big Data Analysis." *International Statistical Review* 87 (1): 177–91. DOI: <https://doi.org/10.1111/insr.12290>.
- Liu, A.-C., S. Scholtus, and T. de Waal. 2023. "Correcting Selection Bias in Big Data by Pseudo-Weighting." *Journal of Survey Statistics and Methodology* 11 (5): 1181–203. DOI: <https://doi.org/10.1093/jssam/smac029>.
- Liu, Y., M. Yuan, P. Li, and C. Wu. 2024. "Statistical Inference with Nonignorable Non-Probability Survey Samples." *ArXiv e-prints*. <https://arxiv.org/abs/2410.02920>.
- Liu, Z., and R. Valliant. 2023. "Investigating an Alternative for Estimation from a Nonprobability Sample: Matching Plus Calibration." *Journal of Official Statistics* 39 (1): 45–78. DOI: <https://doi.org/10.2478/jos-2023-0003>.
- Lohr, S. L., and T. E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statistical Science* 32 (2): 293–312. DOI: <https://doi.org/10.1214/16-STS584>.
- Lothian, J., A. Holmberg, and A. Seyb. 2019. "An Evolutionary Schema for Using 'It-Is-What-It-Is' Data in Official Statistics." *Journal of Official Statistics* 35 (1): 137–65. DOI: <http://dx.doi.org/10.2478/JOS-2019-0007>.

- Mashreghi, Z., D. Haziza, and C. Léger. 2016. "A Survey of Bootstrap Methods in Finite Population Sampling." *Statistical Survey* 10: 1–52. DOI: <https://doi.org/10.1214/16-ss113>.
- Meng, X.-L. 2018. "Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 US Presidential Election." *The Annals of Applied Statistics* 12 (2): 685–726. DOI: <https://doi.org/10.1214/18-aos1161sf>.
- Rao, J. N. K. 2021. "On Making Valid Inferences by Integrating Data from Surveys and Other Sources." *Sankhya B* 83 (1): 242–72. DOI: <https://doi.org/10.1007/s13571-020-00227-w>.
- Rao, J. N. K., and I. Molina. 2015. *Small Area Estimation*. Hoboken, NJ: Wiley.
- Rueda, M., S. Pasadas-del-Amo, B. Rodríguez, L. Castro-Martín, and R. Ferri-García. 2023. "Enhancing Estimation Methods for Integrating Probability and Nonprobability Survey Samples with Machine-Learning Techniques. An Application to a Survey on the Impact of the COVID-19 Pandemic in Spain." *Biometrical Journal* 65 (2): 1–19. DOI: <https://doi.org/10.1002/bimj.202200035>.
- Salvatore, C. 2023. "Inference with Non-Probability Samples and Survey Data Integration: A Science Mapping Study." *Metron* 81 (1): 83–107. DOI: <http://dx.doi.org/10.1007/s40300-023-00243-6>.
- Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer-Verlag. DOI: <https://doi.org/10.1007/978-1-4612-4378-6>.
- Tam, S.-M., and J. K. Kim. 2018. "Big Data Ethics and Selection-Bias: An Official Statistician's Perspective." *Statistical Journal of the IAOS* 34 (4): 577–88. DOI: <https://doi.org/10.3233/sji-170395>.
- Tillé, Y. 2020. *Sampling and Estimation from Finite Populations*. Hoboken, NJ: Wiley. DOI: <https://doi.org/10.1002/9781119071259>.
- Valliant, R. 2009. "Model-Based Prediction of Finite Population Totals." *Handbook of Statistics: Sample Surveys: Inference and Analysis* 29B: 11–31. DOI: [https://doi.org/10.1016/S0169-7161\(09\)00223-5](https://doi.org/10.1016/S0169-7161(09)00223-5).
- Wolter, K. M. 2007. *Introduction to Variance Estimation*. 2nd ed. New York: Springer-Verlag. DOI: <https://doi.org/10.1007/978-0-387-35099-8>.
- Wu, C. 2022. "Statistical Inference with Non-Probability Survey Samples." *Survey Methodology* 48 (2): 283–311.
- Wu, C., and M. E. Thompson. 2020. *Sampling Theory and Practice*. Springer. DOI: <https://doi.org/10.1007/978-3-030-44246-0>.
- Yang, S., and J. K. Kim. 2020. "Statistical Data Integration in Survey Sampling: A Review." *Japanese Journal of Statistics and Data Science* 3: 625–50. DOI: <https://doi.org/10.1007/s42081-020-00093-w>.
- Yang, S., J. K. Kim, and R. Song. 2020. "Doubly Robust Inference When Combining Probability and Non-Probability Samples with High Dimensional Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (2): 445–65. DOI: <https://doi.org/10.1111/rssb.12354>.
- Zhang, L.-C. 2019. "On Valid Descriptive Inference from Non-Probability Sample." *Statistical Theory and Related Fields* 3 (2): 103–13. DOI: <https://doi.org/10.1080/24754269.2019.1666241>.

Received: July 6, 2023

Accepted: March 11, 2025