# Doubly Robust Estimation of the Finite Population Distribution Function Using Nonprobability Samples

Soonpil Kwon , Dongmin Jang , Kyu-Seong Kim *

Posted Date: 15 September 2025

doi: 10.20944/preprints202509.1229.v1

Keywords: nonprobability samples; finite distribution function; quantiles; data integration; doubly robust inference

 doi:10.20944/preprints202509.1229.v1



*Article*

# Doubly Robust Estimation of the Finite Population Distribution Function Using Nonprobability Samples

**Soonpil Kwon [1,2], Dongmin Jang [1] and Kyu-Seong Kim [3,*]**

[1] Department of Statistics and Data Science, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Republic of Korea
[2] Social Statistics Bureau, Statistics Korea, 189 Cheongsa-ro, Seo-gu, Daejeon 35208, Republic of Korea
[3] Department of Statistics, University of Seoul, 163 Seoulsiripdae-ro, Dongdaemun-gu, Seoul 02504, Republic of Korea
* Correspondence: kskim@uos.ac.kr

**Abstract**

In official statistics, quantiles and other distributional indicators are essential for monitoring income distribution and inequality and for producing policy-relevant benchmarks such as medians, deciles, and thresholds. Correspondingly, extensive methods exist for complex survey designs; by contrast, practical tools for nonprobability samples remain limited despite their growing use. We address this gap by developing estimators of the finite population distribution function within a data integration framework that combines nonprobability and probability samples. Specifically, we propose an inverse probability weighted estimator, a regression estimator, and a doubly robust estimator, and we show that the doubly robust estimator is asymptotically unbiased for the finite distribution function when either the propensity score model or the outcome regression model is correctly specified. Building upon the estimated finite distribution function, we derive quantile estimators and construct Woodruff confidence intervals, employing bootstrap variance estimation specifically tailored to this framework. Simulation studies based on a synthetic population and the 2023 Korean Survey of Household Finances and Living Conditions indicate that the proposed estimators perform well in practice, supporting their usefulness for official statistical production.

**Keywords:** nonprobability samples; finite distribution function; quantiles; data integration; doubly robust inference

## 1. Introduction

Probability sampling, first introduced in Neyman's seminal work [1], has long been regarded as the gold standard for finite population inference in survey statistics [2]. In contrast, nonprobability samples were historically employed mainly in observational studies or as supplementary sources for probability sampling. In recent years, however, probability sampling has encountered growing challenges, including rising survey costs, incomplete sampling frames, declining response rates, and increasingly strict data privacy regulations [3,4]. These difficulties have renewed interest in nonprobability samples as a practical alternative for population inference [5,6].

A defining feature of nonprobability samples is their unknown selection mechanism, which, if ignored, can lead to substantial selection bias [7–9]. Moreover, the theoretical foundations for inference with nonprobability samples remain underdeveloped, and standardized criteria for evaluating the reliability of resulting estimates are largely lacking. A prominent strategy to address these limitations is data integration, which leverages information from high-quality probability samples to adjust for the selection bias inherent in nonprobability samples. While early research on data integration focused primarily on combining two probability samples [10,11], more recent work has expanded the scope to include methods that incorporate diverse nonprobability sources, such as online panels and opt-in datasets [12,13].

 doi:10.20944/preprints202509.1229.v1



Within the data integration framework, several approaches have been developed to mitigate selection bias. For example, Elliott and Valliant [13] proposed an approach that models the selection mechanism of the nonprobability sample and applies the inverse of estimated propensity scores as weights. Alternatively, Kim et al. [14] introduced a mass imputation method, which fits an outcome regression model using study variables observed in the nonprobability sample and predicts corresponding values for units in the probability sample. A central limitation of both approaches, however, is that their validity depends entirely on correct model specification.

To overcome these limitations, doubly robust estimation methods have attracted considerable attention, as they yield asymptotically unbiased estimates if at least one of the two underlying models is correctly specified [6,15–17]. Existing applications of doubly robust estimation, however, have largely concentrated on measures of central tendency, such as the population mean, with relatively little attention to distributional estimands—such as finite distribution function and quantiles—that are essential for characterizing the overall shape of a population distribution. While the mean is intuitive and easily interpretable, it is highly sensitive to extreme values and may be inadequate when information about specific regions of the distribution is crucial, for instance in studies of income inequality or health outcomes. By contrast, finite distribution function and quantiles provide information across the entire distribution, enabling more comprehensive analyses [18,19].

In this study, we extend the doubly robust estimator for the population mean proposed by Chen et al. [16] to the estimation of the finite distribution function within the data integration framework. Specifically, we develop three estimators based on auxiliary variables observed in both probability and nonprobability samples: an inverse probability weighted estimator, a regression estimator, and a doubly robust estimator. The proposed doubly robust estimator attains asymptotic unbiasedness for the finite distribution function, provided that either the propensity score model or the outcome regression model is correctly specified.

The remainder of this paper is organized as follows. Section 2 presents the basic setup and assumptions. Section 3 defines the three estimators of the finite distribution function and analyzes their theoretical properties. Section 4 applies these estimators to the estimation of population quantiles and describes the construction of Woodruff confidence intervals. Section 5 evaluates the performance of the proposed methods through simulation studies. Finally, Section 6 concludes with the implications of the study and directions for future research.

## 2. Basic Setup

Consider a finite population $U = \{1, 2, \ldots, N\}$ of size $N$. Each unit $i \in U$ is associated with auxiliary variables $\boldsymbol{x}_i$ and a study variable $y_i$. The parameter of interest in this study is the finite population distribution function, defined as

$$F_y(t) = \frac{1}{N} \sum_{i \in U} I(y_i \leq t), \quad -\infty < t < \infty, \tag{1}$$

where $I(A)$ is the indicator function that equals 1 if the event $A$ is true and 0 otherwise.

Suppose two samples are drawn from the population. The first is a nonprobability sample $S_A$ of size $n_A$, in which both the auxiliary variables $\boldsymbol{x}_i$ and the study variable $y_i$ are observed for each unit in $i \in S_A$. The second is a probability sample $S_B$ of size $n_B$, in which the auxiliary variables $\boldsymbol{x}_i$ are observed together with their inclusion probabilities $\pi_i^B = P(i \in S_B)$ under a given sampling design $p$. In this setup, the nonprobability sample is not representative of the population, whereas the probability sample does not contain observations on the study variable. An integrated approach is therefore required for population inference. This approach relies on common auxiliary variables observed in both samples, which serve as a bridge linking the study variable with the design information. Such a data integration framework is well-established and has been widely applied in prior studies [6,12,13,16,17,20,21]. Following Chen et al. [16], we adopt this framework and assume that the two samples are drawn independently, which simplifies the analysis and enhances theoretical validity.

 doi:10.20944/preprints202509.1229.v1



To utilize nonprobability samples for population inference, we assume that they are drawn under an implicit probabilistic selection mechanism, even though their inclusion probabilities are unknown. This corresponds to the concept of an unknown probability sample discussed by Wu [22]. Under this assumption, the issue of undercoverage is excluded from the scope of this study.

To formalize the selection mechanism, define the indicator variable $R_i$ as

$$R_i = \begin{cases} 1, & \text{if } i \in S_A, \\ 0, & \text{if } i \notin S_A. \end{cases}$$

The conditional expectation of $R_i$ is

$$\pi_i^A = \mathrm{E}_\delta[R_i \mid \boldsymbol{x}_i, y_i] = P(R_i = 1 \mid \boldsymbol{x}_i, y_i), \tag{2}$$

where the subscript $\delta$ denotes the model for the selection mechanism. This probability $\pi_i^A$ is commonly referred to as the propensity score in observational studies [23] and as the participation probability in survey sampling [6,24].

We adopt the concept of an unknown probability sample and impose the following assumptions on the propensity score, as introduced in Chen et al. [16].

**A1** Given the auxiliary variables $\boldsymbol{x}_i$, the study variable $y_i$ and the selection indicator $R_i$ are conditionally independent.

**A2** For all units $i \in U$, $\pi_i^A > 0$.

**A3** Given $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, the selection indicators $R_1, R_2, \ldots, R_N$ are independent.

Assumption **A1** and **A2** together constitute the strong ignorability condition [23]. Under this condition, $\pi_i^A$ in Equation (2) depends only on the auxiliary variables $\boldsymbol{x}_i$.

## 3. Estimators of the Finite Distribution Function

In this section, we introduce three estimators for the finite distribution function in Equation (1) under the data integration framework: the inverse probability weighted (IPW) estimator, the regression (REG) estimator, and the doubly robust (DR) estimator. Their statistical properties are examined within a joint randomization framework consisting of the following three components [17]:

(i) $\delta$: selection mechanism for the nonprobability sample
(ii) $p$: the probability sampling design
(iii) $M$: the regression model for the study variable $y$

Specifically, the IPW estimator is analyzed under the $\delta p$-framework, the REG estimator under the $Mp$-framework, and the DR estimator under either the $\delta p$- or the $Mp$-framework, without specification of which one. All these frameworks incorporate the probability sampling design $p$.

Following Chen et al. [16], we adopt conditions **C1**, **C4**, **C5**, and **C6**, which are redefined as Regularity Conditions **B1**–**B4** in this paper, modified by substituting $I(y_i \le t)$ for $y_i$ and $G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta})$ for $m(\boldsymbol{x}_i, \boldsymbol{\beta})$. To ensure the validity of the Taylor expansion, we also impose the following additional condition

**B5** The distribution function $G(t)$ of the error term is twice differentiable for all $t$.

For asymptotic analysis, we consider a sequence of finite populations indexed by $\nu$, denoted as $U_\nu : \nu = 1, 2, \ldots$ with corresponding samples $S_{A,\nu}$ and $S_{B,\nu}$. As $\nu \to \infty$, the population size $N_\nu$ and the sample sizes $n_{A,\nu}, n_{B,\nu}$ all diverge to infinity. For simplicity, the subscript $\nu$ is omitted hereafter, and asymptotics are expressed in terms of $N \to \infty$.

 doi:10.20944/preprints202509.1229.v1



### 3.1. Inverse Probability Weighted Estimator

The IPW method is widely used to correct for selection bias in nonprobability samples, where the inclusion probability $\pi_i^A$ is assumed to be a function of auxiliary variables $\boldsymbol{x}_i$ and unknown parameters $\boldsymbol{\theta}$ of a participation model

$$\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})$$

Because auxiliary information for the entire population is often unavailable, Chen et al. [16] proposed a pseudo-likelihood approach that incorporates the design weights from a probability sample. The participation model parameters are first estimated by $\hat{\boldsymbol{\theta}}$, which are then used to compute estimated inclusion probabilities $\hat{\pi}_i^A = \pi(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}})$. The corresponding pseudo-weight is $\hat{d}_i^A = 1/\hat{\pi}_i^A$, which is used to construct a Hájek-type estimator of the finite distribution function. This quasi-randomization approach enables valid inference from nonprobability samples [13,25].

For a fixed point $t$, the IPW estimator of the finite distribution function is defined as

$$\hat{F}_{\mathrm{IPW}}(t) = \frac{1}{\hat{N}_A} \sum_{i \in S_A} \hat{d}_i^A I(y_i \le t), \tag{3}$$

where $\hat{N}_A = \sum_{i \in S_A} \hat{d}_i^A$.

**Result 1.** *Under **A1–A3** and **B1–B5**, and if the propensity score model is correctly specified as a logistic regression model, $\hat{F}_{IPW}(t)$ is an asymptotically $\delta p$-unbiased estimator of $F_y(t)$ for fixed point $t$.*

**Proof.** The asymptotic property of the IPW estimator for the population mean $\mu_y$, denoted as $\hat{\mu}_y = \hat{N}_A^{-1} \sum_{i \in S_A} y_i / \hat{\pi}_i^A$, is given by Chen et al. [16] as $\hat{\mu}_y - \mu_y = O_p(n_A^{-1/2})$. The finite distribution function estimator $\hat{F}_{\mathrm{IPW}}(t)$ can be viewed as a plug-in estimator with $y_i$ replaced by $I(y_i \le t)$. Therefore,

$$\hat{F}_{\mathrm{IPW}}(t) - F_y(t) = O_p(n_A^{-1/2}), \quad -\infty < t < \infty$$

Moreover, since both $\hat{F}_{\mathrm{IPW}}(t)$ and $F_y(t)$ lie in $[0,1]$, we have $|\hat{F}_{\mathrm{IPW}}(t) - F_y(t)| \le 1$. Hence,

$$\mathrm{E}_{\delta p}\left[\hat{F}_{\mathrm{IPW}}(t) - F_y(t)\right] = O(n_A^{-1/2})$$

Thus, under the $\delta p$-randomization framework, $\hat{F}_{\mathrm{IPW}}(t)$ is an asymptotically $\delta p$-unbiased estimator of the finite distribution function $F_y(t)$. □

The IPW estimator is asymptotically unbiased under the assumption of strong ignorability, that is, when the selection mechanism is fully explained by the auxiliary variables, a condition also referred to as missing at random (MAR). However, if the propensity score model is misspecified, asymptotic unbiasedness may fail. Moreover, even under correct specification, extreme propensity score values (close to 0 or 1) may yield highly unstable inverse probability weights, inflating the variance of the estimator [26].

### 3.2. Regression Estimator

The REG estimator fits an outcome regression of $y_i$ on $\boldsymbol{x}_i$ using the nonprobability sample $S_A$, and then applies the fitted model to units in the probability sample $S_B$ to estimate the finite distribution function. Because it imputes the study variable for all units in $S_B$, this procedure is commonly referred to as mass imputation. This approach is based on a superpopulation model, in which the finite population $U = \{1, 2, \ldots, N\}$ is assumed to be a random sample generated from the following outcome regression model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{4}$$

where the error terms $\epsilon_i$ follow a normal distribution with $\mathrm{E}_M[\epsilon_i] = 0$ and $\mathrm{Var}_M[\epsilon_i] = \sigma^2$, and are assumed to be independent. The subscript $M$ indicates that the corresponding expectation or

 doi:10.20944/preprints202509.1229.v1



variance is taken under the model (4). By Assumption **A1**, we have $\mathrm{E}_M[y_i \mid \boldsymbol{x}_i, R_i] = \mathrm{E}_M[y_i \mid \boldsymbol{x}_i]$ and $\mathrm{Var}_M[y_i \mid \boldsymbol{x}_i, R_i] = \mathrm{Var}_M[y_i \mid \boldsymbol{x}_i]$, which ensures that the model is valid for the nonprobability sample as well.

A simple model-based approach to estimating the finite distribution function replaces $I(y_i \le t)$ with the plug-in indicator $I(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \le t)$. However, since $\mathrm{E}_M[I(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \le t)] \neq P(y_i < t)$ in general, this naive substitution can lead to bias [18]. To address this, Chambers and Dunstan [27] proposed a residuals-based estimator. Let $G(\cdot)$ denote the distribution function of $\epsilon_i$. The residual-based estimator is then defined as

$$\hat{G}_i(t)_{\mathrm{REG}} = \frac{1}{n_A} \sum_{j \in S_A} I(y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} \le t - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}), \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimator of $\boldsymbol{\beta}$.

Based on Equation (5), the regression estimator of $F_y(t)$ at a fixed point $t$ is given by

$$\hat{F}_{\mathrm{REG}}(t) = \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B \hat{G}_i(t)_{\mathrm{REG}}, \tag{6}$$

where $\hat{N}_B = \sum_{i \in S_B} d_i^B$.

**Result 2.** *Under **A1–A3** and **B1–B5**, $\hat{F}_{REG}(t)$ is an asymptotically $Mp$-unbiased estimator of $F_y(t)$ for fixed point $t$.*

**Proof.** Following the lines of Chambers et al. [28], it can be shown that the result holds under conditions **B1–B5**

$$\mathrm{E}_M\left[\hat{G}_i(t)_{\mathrm{REG}}\right] = G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1}).$$

Using this result, the expectation of the bias can be evaluated as

$$\begin{aligned}
\mathrm{E}\left[\hat{F}_{\mathrm{REG}}(t) - F_y(t)\right] &= \mathrm{E}_p\left[\mathrm{E}_M\left[\hat{F}_{\mathrm{REG}}(t) - F_y(t)\right]\right] \\
&= \mathrm{E}_p\left[\frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B \left\{ G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1}) \right\} - \frac{1}{N} \sum_{i=1}^{N} G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) \right] \\
&\approx \frac{1}{N} \sum_{i=1}^{N} G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \frac{1}{N} \sum_{i=1}^{N} G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1}) \\
&= O(n_A^{-1})
\end{aligned}$$

Therefore, under the $Mp$-randomization framework, $\hat{F}_{\mathrm{REG}}(t)$ is an asymptotically $Mp$-unbiased estimator for the finite distribution function $F_y(t)$. □

When the outcome model is correctly specified, the REG estimator is highly efficient and supports broader use of nonprobability samples. However, if the regression model fails to capture the true distribution, bias may arise and the method becomes sensitive to misspecification. To mitigate this limitation, the next section introduces the DR estimator, which remains asymptotically unbiased provided that either the propensity score model or the outcome regression model is correctly specified.

### *3.3. Doubly Robust Estimator*

The asymptotic unbiasedness of the IPW estimator in Equation (3) and REG estimator in Equation (6) hinges on correct specification of their respective working models. In practice, however, such correctness is difficult to guarantee, motivating procedures that are robust to model misspecification. The DR estimator was introduced to address this issue and has been regarded as a successful approach since Robins et al. [29].

To construct a DR estimator for the finite distribution function, we require an analogue of $\hat{G}_i(t)_{\mathrm{REG}}$ in Equation (5) that estimates the error distribution $G(\cdot)$ and remains valid under the joint randomiza-

 doi:10.20944/preprints202509.1229.v1



tion. Because $\hat{G}_i(t)_{\mathrm{REG}}$ is derived under the $Mp$-framework, it cannot be directly applied when the selection mechanism $\delta$ is operative (i.e., under the $\delta p$-framework). Accordingly, we extend the method of Rao et al. [30] and propose a new estimator of the error distribution that is valid under such joint randomization.

$$\hat{G}_i(t)_{\mathrm{DR}} = \frac{1}{\hat{N}_A} \sum_{j \in S_A} \hat{d}_j^A I(y_j - \boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}} \le t - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \tag{7}$$

Based on $\hat{G}_i(t)_{\mathrm{DR}}$ defined in Equation (7), the DR estimator of the finite distribution function $F_y(t)$ at a fixed point $t$ is then given by

$$\hat{F}_{\mathrm{DR}}(t) = \frac{1}{\hat{N}_A} \sum_{i \in S_A} \hat{d}_i^A \{ I(y_i \le t) - \hat{G}_i(t)_{\mathrm{DR}} \} + \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B \hat{G}_i(t)_{\mathrm{DR}} \tag{8}$$

**Result 3.** *Under regularity conditions **A1–A3** and **B1–B5**, and if at least one of the propensity score model or the outcome regression model is correctly specified, $F_{DR}(t)$ is an asymptotically unbiased estimator of $F_y(t)$ at a fixed point $t$ under the $\delta p$- or $Mp$-framework.*

**Proof.** (i) When the propensity score model is correctly specified

The doubly robust estimator can be rewritten as

$$\hat{F}_{\mathrm{DR}}(t) = \hat{F}_{\mathrm{IPW}}(t) - \frac{1}{\hat{N}_A} \sum_{i \in S_A} \hat{d}_i^A \hat{G}_i(t)_{\mathrm{DR}} + \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B \hat{G}_i(t)_{\mathrm{DR}}$$

The second and third terms on the right-hand side are Hájek estimators of $N^{-1} \sum_{i=1}^N \hat{G}_i(t)_{\mathrm{DR}}$ based on the nonprobability sample $S_A$ and the probability sample $S_B$, respectively, and hence cancel out asymptotically. Given the asymptotic $\delta p$-unbiasedness of $\hat{F}_{\mathrm{IPW}}(t)$, $\hat{F}_{\mathrm{DR}}(t)$ is also asymptotically $\delta p$-unbiased.

(ii) When the outcome regression model is correctly specified

Similarly to the proof for the REG estimator, we have

$$\mathrm{E}_M[\hat{G}_i(t)_{\mathrm{DR}}] = G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1})$$

Using this, the expected bias is

$$\begin{aligned}
&\mathrm{E}[\hat{F}_{\mathrm{DR}}(t) - F_y(t)] \\
&= \mathrm{E}_p \left[ \frac{1}{\hat{N}_A} \sum_{i \in S_A} \hat{d}_i^A \mathrm{E}_M[I(y_i \le t) - \hat{G}_i(t)_{\mathrm{DR}}] + \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B \mathrm{E}_M[\hat{G}_i(t)_{\mathrm{DR}}] - \frac{1}{N} \sum_{i \in U} \mathrm{E}_M[I(y_i \le t)] \right] \\
&= \mathrm{E}_p \left[ \frac{1}{\hat{N}_B} \sum_{i \in S_B} d_i^B G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \frac{1}{N} \sum_{i=1}^N G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1}) \right] \\
&\approx \frac{1}{N} \sum_{i=1}^N G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) - \frac{1}{N} \sum_{i=1}^N G(t - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + O(n_A^{-1}) \\
&= O(n_A^{-1})
\end{aligned}$$

Therefore, under the $Mp$-framework, $\hat{F}_{\mathrm{DR}}(t)$ is an asymptotically unbiased estimator of the finite distribution function $F_y(t)$. □

The asymptotic unbiasedness of the DR estimator requires the estimated coefficients to satisfy probability-limit conditions; specifically, for the propensity score parameters $\hat{\boldsymbol{\theta}}$ and the outcome regression parameters $\hat{\boldsymbol{\beta}}$, there exist fixed vectors $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$ such that $p\lim \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*$ and $p\lim \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^*$. If the propensity score model is correctly specified, then $\boldsymbol{\theta}^* = \boldsymbol{\theta}$, and if the outcome regression model

 doi:10.20944/preprints202509.1229.v1



is correctly specified, then $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. Under misspecification, by contrast, these probability limits need not coincide with the true parameters, and the limiting value itself does not have a meaningful interpretation.

## 4. Quantile Estimation

An important application of the finite distribution function estimators is the estimation of population quantiles, defined as

$$\xi_q = \inf\{t; F_y(t) \geq q\}$$

Quantiles provide informative summaries of distributional features such as central tendency, spread, and asymmetry, and they are useful for assessing the presence of outliers. Because estimators of the finite distribution function are typically step functions, linear interpolation is employed to obtain a unique estimate of the $q$th quantile [30–32]. The quantile estimator $\hat{\xi}_q$ is expressed as

$$\hat{\xi}_q = a + \frac{q - \hat{F}(a)}{\hat{F}(b) - \hat{F}(a)}(b - a),$$

where $a = \max\{t; \hat{F}(t) \leq q\}$ and $b = \min\{t; \hat{F}(t) \geq q\}$.

A widely used method for constructing a confidence interval (CI) for a quantile estimator was proposed by Woodruff [31]. The key idea is to first obtain a CI for the estimated finite distribution function and then invert this interval to derive a CI for the quantile. The resulting $100(1-\alpha)\%$ CI is given by

$$\hat{\xi}_q^L = \inf\left\{t; \hat{F}(t) \geq q - z_{1-\alpha/2}\sqrt{\hat{V}[\hat{F}(\hat{\xi}_q)]}\right\},$$
$$\hat{\xi}_q^U = \inf\left\{t; \hat{F}(t) \geq q + z_{1-\alpha/2}\sqrt{\hat{V}[\hat{F}(\hat{\xi}_q)]}\right\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and $\hat{V}[\hat{F}(\hat{\xi}_q)]$ denotes the estimated variance of $\hat{F}(t)$ evaluated at $\hat{\xi}_q$. Sitter and Wu [33] provided empirical evidence that the Woodruff method attains approximately correct coverage even for extreme quantiles (large or small $q$).

## 5. Simulation Studies

To evaluate the performance of the proposed finite distribution function estimators, $\hat{F}_{\text{IPW}}(t)$, $\hat{F}_{\text{REG}}(t)$, and $\hat{F}_{\text{DR}}(t)$, we conducted simulation studies based on two populations: A synthetic finite population from Chen et al. [16], and the 2023 Korean Survey of Household Finances and Living Conditions.

The variances of the finite distribution function estimators were obtained via a bootstrap procedure following Chen et al. [16]:

1. From the nonprobability sample $S_A$ and the probability sample $S_B$, draw bootstrap samples $S_A^{(j)}$ and $S_B^{(j)}$ of sizes $n_A$ and $n_B$, respectively, by simple random sampling with replacement, for $J = 1,000$ replicates.
2. For each bootstrap replicate, compute $\hat{F}^{(j)}(t)$.
3. Using $\{F^{(j)}(t)\}_{j=1}^J$ calculate the bootstrap variance estimator $v_{\text{BT}}$.

Performance was then assessed over $R = 3,000$ simulation replications using percentage relative bias (%RB) and relative root mean squared error (RRMSE), where

$$\%\text{RB} = \frac{1}{R}\sum_{r=1}^{R}\frac{\hat{\theta}^{(r)} - \theta}{\theta} \times 100, \quad \text{RRMSE} = \frac{1}{\theta}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)} - \theta\right)^2},$$

 doi:10.20944/preprints202509.1229.v1



with $\hat{\theta}^{(r)}$ denoting the estimate from replication $r$ and $\theta$ the target parameter. For the finite distribution function, the bootstrap variance, and quantiles, the corresponding choices were

- The finite distribution function: $\hat{\theta}^{(r)} = \hat{F}^{(r)}(t), \theta = F_y(t)$
- Bootstrap variance: $\hat{\theta}^{(r)} = v_{\mathrm{BT}}^{(r)}, \theta = V$
- Quantile: $\hat{\theta}^{(r)} = \hat{\xi}_q^{(r)}, \theta = \xi_q$

where $V$ denotes the simulation-based variance of $\hat{F}(t)$ computed from 10,000 replications.

The coverage probability of the confidence interval based on the bootstrap variance (%CP$_v$) was evaluated as

$$\%\mathrm{CP}_v = \frac{1}{R}\sum_{r=1}^{R} I\left(|\hat{F}^{(r)}(t) - F_y(t)| \le z_{1-\alpha/2}\sqrt{v_{\mathrm{BT}}^{(r)}}\right).$$

The performance of the Woodruff confidence interval was assessed by its coverage probability (%CP$\xi$), lower error rate (%L), and upper error rate (%U):

$$\%\mathrm{CP}_\xi = \frac{1}{R}\sum_{r=1}^{R} I\left(\xi_q^{L(r)} < \xi_q < \xi_q^{U(r)}\right),$$

$$\%\mathrm{L} = \frac{1}{R}\sum_{r=1}^{R} I\left(\xi_q < \xi_q^{L(r)}\right),$$

$$\%\mathrm{U} = \frac{1}{R}\sum_{r=1}^{R} I\left(\xi_q^{U(r)} < \xi_q\right),$$

where $\xi_q^{L(r)}$ and $\xi_q^{U(r)}$ denote, respectively, the lower and upper Woodruff CI bounds for the $q$th quantile in replication $r$.

### *5.1. Study 1*

Following the simulation design of Chen et al. [16], we generated a finite population of size $N = 20,000$. The study variable $y$ and auxiliary variables $\boldsymbol{x}$ are generated from

$$y_i = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \sigma\epsilon_i, \quad i = 1, 2, \ldots, N,$$

where $(x_{1i}, x_{2i}, x_{3i}, x_{4i})$ follow the design in Chen et al. [16], and the error terms $\epsilon_i \sim N(0,1)$. The parameter $\sigma$ is chosen such that the correlation coefficient $\rho$ between $y$ and the linear predictor $\boldsymbol{x}^\top\boldsymbol{\beta}$ equals 0.5.

We consider four model specification scenarios

- TT: Both $\delta$ and $M$ are correctly specified.
- TF: $\delta$ is correctly specified, but $M$ is misspecified, with $x_{4i}$ omitted from the model.
- FT: $M$ is correctly specified, but $\delta$ is misspecified, with $x_{4i}$ omitted from the model.
- FF: Both models are misspecified, with $x_{4i}$ omitted in each model.

The analysis uses a nonprobability sample $S_A$ of size $n_A = 500$ and a probability sample $S_B$ of size $n_B = 1,000$. Table 1 reports %RB and RRMSE for the proposed finite distribution function estimators. Under TT, all estimators exhibit low bias and error, indicating stable performance. Under TF and FT, the DR estimator attains lower bias and error than the alternatives, highlighting the advantages of the doubly robust property. By contrast, under FF, performance deteriorates substantially for all estimators.

 doi:10.20944/preprints202509.1229.v1



**Table 1.** %RB and RRMSE of the Finite Distribution Function Estimators (Study 1).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE |
| TT | $\hat{F}_{\text{IPW}}(t)$ | 0.60 | 0.10 | 0.30 | 0.06 | 0.23 | 0.03 |
| | $\hat{F}_{\text{REG}}(t)$ | 0.22 | 0.08 | 0.01 | 0.04 | -0.28 | 0.02 |
| | $\hat{F}_{\text{DR}}(t)$ | 0.24 | 0.10 | 0.01 | 0.05 | 0.06 | 0.03 |
| TF | $\hat{F}_{\text{IPW}}(t)$ | 0.60 | 0.10 | 0.30 | 0.06 | 0.23 | 0.03 |
| | $\hat{F}_{\text{REG}}(t)$ | -30.27 | 0.31 | -24.23 | 0.25 | -17.02 | 0.17 |
| | $\hat{F}_{\text{DR}}(t)$ | 0.53 | 0.10 | 0.25 | 0.06 | 0.20 | 0.03 |
| FT | $\hat{F}_{\text{IPW}}(t)$ | -30.06 | 0.31 | -23.91 | 0.24 | -16.88 | 0.17 |
| | $\hat{F}_{\text{REG}}(t)$ | 0.22 | 0.08 | 0.01 | 0.04 | -0.28 | 0.02 |
| | $\hat{F}_{\text{DR}}(t)$ | 0.33 | 0.09 | 0.31 | 0.05 | 0.09 | 0.03 |
| FF | $\hat{F}_{\text{IPW}}(t)$ | -30.06 | 0.31 | -23.91 | 0.24 | -16.88 | 0.17 |
| | $\hat{F}_{\text{REG}}(t)$ | -30.27 | 0.31 | -24.23 | 0.25 | -17.02 | 0.17 |
| | $\hat{F}_{\text{DR}}(t)$ | -30.01 | 0.31 | -23.88 | 0.24 | -16.86 | 0.17 |

Table 2 compares the bootstrap variance estimators in terms of %RB and $\%\text{CP}_v$. Under TT, all variance estimators perform satisfactorily. Under TF and FT, despite model misspecification, the variance estimator associated with the DR method retains low bias and an $\%\text{CP}_v$ close to 95%, indicating stable reliability and accuracy. Conversely, under FF, coverage performance deteriorates markedly across all methods.

**Table 2.** %RB and $\%\text{CP}_v$ of Bootstrap Variance Estimators (Study 1).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | $\%\text{CP}_v$ | %RB | $\%\text{CP}_v$ | %RB | $\%\text{CP}_v$ |
| TT | $v_{\text{IPW, BT}}$ | 6.11 | 95.50 | 6.98 | 95.10 | 7.16 | 95.40 |
| | $v_{\text{REG, BT}}$ | 3.07 | 95.00 | 3.50 | 94.90 | 5.51 | 96.40 |
| | $v_{\text{DR, BT}}$ | 2.39 | 95.10 | 3.45 | 95.50 | 5.02 | 96.00 |
| TF | $v_{\text{IPW, BT}}$ | 6.11 | 95.50 | 6.98 | 95.10 | 7.16 | 95.40 |
| | $v_{\text{REG, BT}}$ | 6.01 | 0.60 | 5.36 | 0.00 | 9.52 | 0.00 |
| | $v_{\text{DR, BT}}$ | 5.04 | 95.90 | 5.42 | 95.00 | 5.05 | 95.60 |
| FT | $v_{\text{IPW, BT}}$ | 3.01 | 2.20 | 4.01 | 0.00 | 8.60 | 0.00 |
| | $v_{\text{REG, BT}}$ | 3.07 | 95.00 | 3.50 | 94.90 | 5.51 | 96.40 |
| | $v_{\text{DR, BT}}$ | 1.78 | 95.50 | 3.26 | 95.80 | 6.33 | 96.00 |
| FF | $v_{\text{IPW, BT}}$ | 3.01 | 2.20 | 4.01 | 0.00 | 8.60 | 0.00 |
| | $v_{\text{REG, BT}}$ | 6.01 | 0.60 | 5.36 | 0.00 | 9.52 | 0.00 |
| | $v_{\text{DR, BT}}$ | 2.98 | 2.30 | 3.78 | 0.00 | 8.47 | 0.00 |

Table 3 summarizes the results for the quantile estimators. Mirroring the findings for the finite distribution function estimators, all methods perform well under the TT scenario. Under TF and FT, the DR-based quantiles remain stable, confirming the robustness of the doubly robust approach. By contrast, under FF, overall estimation accuracy deteriorates.

 doi:10.20944/preprints202509.1229.v1



**Table 3.** %RB and RRMSE of Quantile Estimators (Study 1).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE |
| TT | $\hat{\xi}_q^{\mathrm{IPW}}$ | -1.57 | 0.18 | -0.56 | 0.06 | -0.33 | 0.04 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | -0.31 | 0.13 | -0.04 | 0.05 | 0.33 | 0.03 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | -0.94 | 0.17 | -0.21 | 0.06 | -0.21 | 0.04 |
| TF | $\hat{\xi}_q^{\mathrm{IPW}}$ | -1.57 | 0.18 | -0.56 | 0.06 | -0.33 | 0.04 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 60.70 | 0.62 | 30.24 | 0.31 | 22.65 | 0.23 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | -1.40 | 0.18 | -0.45 | 0.07 | -0.35 | 0.05 |
| FT | $\hat{\xi}_q^{\mathrm{IPW}}$ | 60.60 | 0.63 | 29.49 | 0.30 | 22.84 | 0.23 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | -0.31 | 0.13 | -0.04 | 0.05 | 0.33 | 0.03 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | -0.97 | 0.15 | -0.46 | 0.05 | -0.29 | 0.04 |
| FF | $\hat{\xi}_q^{\mathrm{IPW}}$ | 60.60 | 0.63 | 29.49 | 0.30 | 22.84 | 0.23 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 60.70 | 0.62 | 30.24 | 0.31 | 22.65 | 0.23 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | 60.48 | 0.62 | 29.48 | 0.30 | 22.81 | 0.23 |

Table 4 reports the Woodruff CI results for the quantile estimators, including %CP$\xi$, %L, and %U.Consistent with previous findings, all methods perform well under the TT scenario. Under TF and FT, the DR-based intervals maintain %CP$\xi$ close to the nominal 95% with balanced tail errors, indicating high reliability. By contrast, under FF, coverage deteriorates substantially across methods. %CP$\xi$ falls below the nominal level and both tail error rates increase, signaling degraded interval performance.

**Table 4.** %CP$_{\xi}$, %L, and %U of Woodruff Confidence Intervals (Study 1).

| | | $t_1 = \xi_{0.25}$ | | | $t_2 = \xi_{0.50}$ | | | $t_3 = \xi_{0.75}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %CP$_{\xi}$ | %L | %U | %CP$_{\xi}$ | %L | %U | %CP$_{\xi}$ | %L | %U |
| TT | $\hat{\xi}_q^{\mathrm{IPW}}$ | 95.70 | 1.83 | 2.47 | 95.37 | 2.20 | 2.43 | 96.37 | 1.73 | 1.90 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 94.43 | 3.30 | 2.27 | 95.07 | 3.23 | 1.70 | 94.70 | 4.27 | 1.03 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | 94.53 | 2.47 | 3.00 | 94.93 | 2.43 | 2.63 | 96.00 | 2.10 | 1.90 |
| TF | $\hat{\xi}_q^{\mathrm{IPW}}$ | 95.70 | 1.83 | 2.47 | 95.37 | 2.20 | 2.43 | 96.37 | 1.73 | 1.90 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 0.57 | 99.43 | 0.00 | 0.03 | 99.97 | 0.00 | 0.03 | 99.97 | 0.00 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | 95.10 | 2.10 | 2.80 | 95.03 | 2.47 | 2.50 | 95.97 | 1.87 | 2.17 |
| FT | $\hat{\xi}_q^{\mathrm{IPW}}$ | 2.53 | 97.47 | 0.00 | 0.03 | 99.97 | 0.00 | 0.03 | 99.97 | 0.00 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 94.43 | 3.30 | 2.27 | 95.07 | 3.23 | 1.70 | 94.70 | 4.27 | 1.03 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | 94.60 | 2.40 | 3.00 | 95.00 | 2.23 | 2.77 | 95.50 | 2.13 | 2.37 |
| FF | $\hat{\xi}_q^{\mathrm{IPW}}$ | 2.53 | 97.47 | 0.00 | 0.03 | 99.97 | 0.00 | 0.03 | 99.97 | 0.00 |
| | $\hat{\xi}_q^{\mathrm{REG}}$ | 0.57 | 99.43 | 0.00 | 0.03 | 99.97 | 0.00 | 0.03 | 99.97 | 0.00 |
| | $\hat{\xi}_q^{\mathrm{DR}}$ | 2.70 | 97.30 | 0.00 | 0.03 | 99.97 | 0.00 | 0.03 | 99.97 | 0.00 |

### *5.2. Study 2*

In the second simulation study, we treat the 2023 Korean Survey of Household Finances and Living Conditions (SHFLC; $(N = 16,730)$) as the finite population and repeatedly draw subsamples from it. Table 5 summarizes the key variables used in the experiment and their definitions.

 doi:10.20944/preprints202509.1229.v1



**Table 5.** Variables and Definitions from the Korean Survey of Household Finances and Living Conditions (2023).

| Variable | Description |
|---|---|
| INCOME | Current income |
| EDU | Educational attainment |
| GEO | Metropolitan status: In metropolitan area (1), Not in metropolitan area (2) |
| SNG | One-person household: Yes (1), No (2) |
| APT | Residence in an apartment: Yes (1), No (2) |
| SIZE | Size of net Floor Area: Classified into 4 groups by size |
| HOME | Housing types |
| DEBT | Any debt held by the household: Yes (1), No (2) |
| EXP1 | Consumption expenditure |
| EXP2 | Non-consumption expenditure |

The nonprobability sample $S_A$ was generated to mimic structures commonly observed in practice. The propensity score model was specified as a logistic regression,

$$\log\left\{\frac{\pi_i^A}{1-\pi_i^A}\right\} = \zeta_0 + \zeta_1 \texttt{EDU} + \zeta_2 \texttt{SNG} + \zeta_3 \texttt{APT} + \zeta_4 \texttt{DEBT}, \quad i = 1, \ldots, N,$$

where $\zeta_0$ was chosen so that $\sum_{i=1}^{N} \pi_i^A = n_A$. Under this design, households with higher educational attainment of the household head, non-single households, apartment residents, and households without debt were more likely to be included in $S_A$. The nonprobability sample $S_A$ was then selected by Poisson sampling with inclusion probabilities $\pi_i^A$.

The probability sample $S_B$ was stratified into nine strata defined by `GEO`, `HOME`, and `SIZE`. A mixed allocation scheme—combining Neyman and proportional allocation—was used to determine stratum specific sample sizes, followed by simple random sampling without replacement within each stratum. The sample sizes were set to $n_A = 500$ and $n_B = 1000$.

The study variable of interest was current income (`INCOME`). Because the true outcome model was unknown, we included `EXP1` and `EXP2`- the covariates with comparatively strong explanatory power- as regressors in the working model. This setup allows us to assess the impact of model misspecification on estimation performance and to isolate efficiency gains attributable to the DR estimator. We consider two scenarios regarding the propensity score model:

- A: correctly specified propensity score model.
- B: misspecified propensity score model (excluding `SNG` and `DEBT`).

Table 6 reports the results for the distribution–function estimators. Overall, the REG estimator performs reasonably well, although its bias and error are somewhat larger at lower quantiles than at middle and upper quantiles, likely reflecting the limited explanatory power of the auxiliary variables and the possible over-representation of high-income households. Under Scenario A, the IPW estimator and the DR estimator both exhibit low bias and error, confirming the effectiveness of propensity score adjustment. Under Scenario B, REG estimator is the most stable, while the DR estimator inherits some bias from the misspecified IPW component and thus loses efficiency. In summary, when the propensity score model is correctly specified, the IPW estimator, the REG estimator, and the DR estimator all yield stable results. However, when the propensity-score model is misspecified, only the REG estimator and the DR estimator perform well, with the REG estimator performing best. These findings highlight that the choice of estimator may critically depend on the availability and explanatory power of the auxiliary variables.

 doi:10.20944/preprints202509.1229.v1



**Table 6.** %RB and RRMSE of the Finite Distribution Function Estimators (Study 2).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE |
| A | $\hat{F}_{\text{IPW}}(t)$ | 0.11 | 0.07 | 0.02 | 0.04 | 0.05 | 0.03 |
| | $\hat{F}_{\text{REG}}(t)$ | -7.80 | 0.09 | -2.60 | 0.04 | 0.55 | 0.02 |
| | $\hat{F}_{\text{DR}}(t)$ | 0.19 | 0.06 | 0.10 | 0.04 | 0.15 | 0.02 |
| B | $\hat{F}_{\text{IPW}}(t)$ | 15.91 | 0.18 | 9.19 | 0.10 | 3.67 | 0.04 |
| | $\hat{F}_{\text{REG}}(t)$ | -7.80 | 0.09 | -2.60 | 0.04 | 0.55 | 0.02 |
| | $\hat{F}_{\text{DR}}(t)$ | 6.48 | 0.09 | 3.14 | 0.05 | 1.08 | 0.02 |

Table 7 compares the bootstrap variance estimators in terms of %RB and%$\text{CP}_v$. Consistent with the findings for the finite distribution function estimators, the REG estimator shows degraded variance performance at lower quantiles. The IPW estimator maintains coverage close to 95% %$\text{CP}_v$ under Scenario A, but its %$\text{CP}_v$ declined markedly under Scenario B. The DR estimator achieves both low bias and stable %$\text{CP}_v$ across scenarios, indicating reliable variance estimation.

**Table 7.** %RB and %$\text{CP}_v$ of Bootstrap Variance Estimators (Study 2).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | %$\text{CP}_v$ | %RB | %$\text{CP}_v$ | %RB | %$\text{CP}_v$ |
| A | $v_{\text{IPW, BT}}$ | 10.03 | 96.10 | 11.95 | 96.07 | 7.61 | 95.67 |
| | $v_{\text{REG, BT}}$ | 16.34 | 62.37 | 25.00 | 88.10 | 17.29 | 94.53 |
| | $v_{\text{DR, BT}}$ | 14.67 | 96.67 | 16.86 | 96.67 | 12.18 | 95.23 |
| B | $v_{\text{IPW, BT}}$ | 7.40 | 47.20 | 9.65 | 42.50 | 7.46 | 66.67 |
| | $v_{\text{REG, BT}}$ | 16.34 | 62.37 | 25.00 | 88.10 | 17.29 | 94.53 |
| | $v_{\text{DR, BT}}$ | 13.26 | 84.90 | 16.73 | 88.10 | 12.47 | 92.57 |

Table 8 compares the quantile estimators in terms of %RB and RRMSE. The REG estimator shows substantial bias at lower quantiles, whereas the IPW estimator performs well under Scenario A but deteriorates under Scenario B. The DR estimator maintains moderate bias and error across both scenarios, yielding comparatively stable performance overall.

**Table 8.** %RB and RRMSE of Quantile Estimators (Study 2).

| | | $t_1 = \xi_{0.25}$ | | $t_2 = \xi_{0.50}$ | | $t_3 = \xi_{0.75}$ | |
|---|---|---|---|---|---|---|---|
| Scenario | Estimator | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE |
| A | $\hat{\xi}_q^{\text{IPW}}$ | -0.18 | 0.06 | -0.14 | 0.05 | -0.28 | 0.04 |
| | $\hat{\xi}_q^{\text{REG}}$ | 7.57 | 0.09 | 2.96 | 0.04 | -0.82 | 0.03 |
| | $\hat{\xi}_q^{\text{DR}}$ | -0.35 | 0.06 | -0.23 | 0.04 | -0.47 | 0.04 |
| B | $\hat{\xi}_q^{\text{IPW}}$ | -14.06 | 0.15 | -10.80 | 0.12 | -6.94 | 0.08 |
| | $\hat{\xi}_q^{\text{REG}}$ | 7.57 | 0.09 | 2.96 | 0.04 | -0.82 | 0.03 |
| | $\hat{\xi}_q^{\text{DR}}$ | -6.32 | 0.08 | -3.91 | 0.06 | -2.12 | 0.04 |

Table 9 reports results for the Woodruff confidence intervals of the quantile estimators-%$\text{CP}_{\xi}$, %L, and %U. The IPW estimator attains %$\text{CP}_{\xi}$ close to the nominal 95% under Scenario A, but coverage drops sharply under Scenario B, accompanied by an upward bias in %U, indicating sensitivity to propensity score misspecification. The REG estimator performs well at the middle and upper quantiles, but shows increased %L at lower quantiles. The DR estimator maintains stable %$\text{CP}_{\xi}$ across scenarios, with only a slight upward bias in %U under Scenario B.

 doi:10.20944/preprints202509.1229.v1



**Table 9.** $\%CP_{\hat{\xi}}$, %L and %U of Woodruff Confidence Intervals (Study 2).

| | | $t_1 = \xi_{0.25}$ | | | $t_2 = \xi_{0.50}$ | | | $t_3 = \xi_{0.75}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Scenario | Estimator | $\%CP_{\hat{\xi}}$ | %L | %U | $\%CP_{\hat{\xi}}$ | %L | %U | $\%CP_{\hat{\xi}}$ | %L | %U |
| A | $\hat{\xi}_q^{IPW}$ | 96.43 | 1.67 | 1.90 | 96.33 | 1.47 | 2.20 | 95.87 | 2.13 | 2.00 |
| | $\hat{\xi}_q^{REG}$ | 58.43 | 41.57 | 0.00 | 83.67 | 16.23 | 0.10 | 95.67 | 1.53 | 2.80 |
| | $\hat{\xi}_q^{DR}$ | 96.87 | 1.60 | 1.53 | 96.90 | 1.33 | 1.77 | 95.70 | 1.87 | 2.43 |
| B | $\hat{\xi}_q^{IPW}$ | 44.57 | 0.00 | 55.43 | 43.03 | 0.00 | 56.97 | 71.57 | 0.03 | 28.40 |
| | $\hat{\xi}_q^{REG}$ | 58.43 | 41.57 | 0.00 | 83.67 | 16.23 | 0.10 | 95.67 | 1.53 | 2.80 |
| | $\hat{\xi}_q^{DR}$ | 83.60 | 0.00 | 16.40 | 88.70 | 0.03 | 11.27 | 94.37 | 0.47 | 5.17 |

## 6. Conclusions

This study proposed three estimators—Inverse Probability Weighting (IPW), Regression-based estimation (REG), and Doubly Robust estimation (DR)—for reliable estimation of the finite population distribution function and quantiles within a data integration framework that combines probability and nonprobability samples. We examined both theoretical properties and empirical performance. In particular, the DR estimator offers a practical advantage: it retains asymptotic unbiasedness for the finite distribution function provided that either the propensity score model or the outcome regression model is correctly specified, thereby affording robustness to the model misspecification that frequently arises in applied survey settings. Building on this theoretical foundation, we conducted simulation studies using two populations: the synthetic population of [16] and the 2023 Korean Survey of Household Finances and Living Conditions. Across various evaluation metrics, the DR-based procedures showed robust performance, maintaining low relative bias, stable relative root mean squared error, and coverage probabilities close to 95% even when one of the models was misspecified. Notably, DR outperformed IPW and REG when the regression model was inaccurate or the propensity score model was partially misspecified, while also yielding balanced results in the presence of over-representation of high-income households and in lower quantile regions. Furthermore, the composition of auxiliary variables was found to be crucial for estimation performance. Inclusion of covariates with strong explanatory power improved the performance of REG and DR, whereas limited auxiliary information led to increased bias in certain cases. This underscores the importance of selecting appropriate auxiliary variables at the stages of survey design and data integration. Overall, these findings demonstrate that the proposed methods can mitigate the limitations of nonprobability samples and highlight their potential applicability in data environments such as online panel surveys and web-based sources where representativeness is often difficult to achieve.

The main contributions of this study can be summarized in two aspects. First, unlike previous doubly robust (DR) methods that have primarily focused on mean estimation, we extended the approach to the estimation of finite population distribution functions and quantiles. This extension enables more precise and flexible analyses in domains where distributional characteristics such as income, consumption, and health are of central importance. Second, the proposed method enhances the utility of nonprobability samples while being naturally integrated into the framework of probability-based inference, thereby providing an analytical framework well suited for modern survey environments where multiple data sources coexist. Nevertheless, several limitations remain. First, the asymptotic unbiasedness of the DR estimator requires that either the propensity score model or the regression model satisfies certain regularity conditions. When sample sizes are small or the distribution of propensity scores is highly imbalanced, the estimation may become unstable. Second, methodologies for variance estimation in data integration settings are not yet fully established. Conventional bootstrap procedures may overestimate variance, indicating the need for refined theoretical approaches. Third, the present study was conducted under the Missing at Random (MAR) assumption. However, in practice, situations of Not Missing at Random (NMAR) and structural undercoverage occur frequently, highlighting the necessity of developing estimation procedures and diagnostic tools that can address such issues. Future research directions include nonparametric or semiparametric propensity score

 doi:10.20944/preprints202509.1229.v1



estimation, integration of high-dimensional auxiliary information through machine learning methods, and applications to a wider range of empirical data sources. Methodological advances along these lines will enable the production of reliable statistics that can accommodate the complexities of real-world survey environments, thereby contributing to evidence-based policymaking using public data.

**Author Contributions:** Conceptualization and methodology, Kwon and Kim; Software and data curation, Jang; Writing—original draft, Kwon; Writing—review & editing, Kwon, Jang, and Kim; Supervision and funding acquisition, Kim.

**Funding:** This research was supported by the National Research Foundation of Korea (NRF), Grant No. RS-2022-NR068754.

**Data Availability Statement:** The Korean Survey of Household Finances and Living Conditions (SHFLC 2023) is available as public-use data from the Microdata Integrated Service (MDIS) of Statistics Korea (KOSTAT). Derived, de-identified analysis outputs used in this study are provided in the repository.

**Acknowledgments:**

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| IPW | Inverse Probability Weighting |
| REG | Regression |
| DR | Doubly Robust |
| MAR | Missing at Random |
| NMAR | Not Missing at Random |

## References

1. Neyman, J. On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society* **1934**, *97*, 558–625.
2. Kim, J.K. A gentle introduction to data integration in survey sampling. *The Survey Statistician* **2022**.
3. Baker, R.; Brick, J.M.; Bates, N.A.; Battaglia, M.; Couper, M.P.; Dever, J.A.; Gile, K.J.; Tourangeau, R. Summary report of the AAPOR task force on non-probability sampling. *Journal of survey statistics and methodology* **2013**, *1*, 90–143.
4. Keiding, N.; Louis, T.A. Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society Series A: Statistics in Society* **2016**, *179*, 319–376.
5. Rancourt, E. Admin-First as a statistical paradigm for Canadian official statistics: Meaning, challenges and opportunities. In Proceedings of the Proceedings of Statistics Canada 2018 International Methodology Symposium, 2018.
6. Beaumont, J.F. Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* **2020**, *46*, 1–29.
7. Harms, T.; Duchesne, P. On calibration estimation for quantiles. *Survey methodology* **2006**, *32*, 37.
8. Meng, X.L. 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* **2018**, *12*, 685–726.
9. Bethlehem, J. Selection bias in web surveys. *International statistical review* **2010**, *78*, 161–188.
10. Wu, C. Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics* **2004**, *32*, 15–26.
11. Kim, J.K.; Rao, J.N. Combining data from two independent surveys: a model-assisted approach. *Biometrika* **2012**, *99*, 85–100.
12. Rivers, D. Sampling for web surveys. In Proceedings of the Joint Statistical Meetings. American Statistical Association Alexandria, VA, 2007, Vol. 4, p. 1320.
13. Elliott, M.R.; Valliant, R. Inference for nonprobability samples. *Statistical Science* **2017**.
14. Kim, J.K.; Park, S.; Chen, Y.; Wu, C. Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society* **2021**, *184*, 941–963.

 doi:10.20944/preprints202509.1229.v1




15. Kim, J.K.; Haziza, D. Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **2014**, *24*, 375–394.
16. Chen, Y.; Li, P.; Wu, C. Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* **2020**, *115*, 2011–2021.
17. Wu, C. Statistical inference with non-probability survey samples. *Survey Methodology* **2022**, *48*, 283–311.
18. Valliant, R.; Dorfman, A.H.; Royall, R.M. *Finite population sampling and inference: a prediction approach*; Wiley New York, 2000.
19. Särndal, C.E.; Swensson, B.; Wretman, J. *Model assisted survey sampling*; Springer Science & Business Media, 2003.
20. Vavreck, L.; Rivers, D. The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties* **2008**, *18*, 355–366.
21. Lee, S.; Valliant, R. Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research* **2009**, *37*, 319–343.
22. Wu, C. Author's response to comments on" Statistical inference with non-probability survey samples", 2022.
23. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.
24. Rao, J. On making valid inferences by integrating data from surveys and other sources. *Sankhya B* **2021**, *83*, 242–272.
25. Kott, P.S. A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association* **1994**, *89*, 693–696.
26. Kang, J.D.; Schafer, J.L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data **2007**.
27. Chambers, R.L.; Dunstan, R. Estimating distribution functions from survey data. *Biometrika* **1986**, *73*, 597–604.
28. Chambers, R.; Dorfman, A.H.; Hall, P. Properties of estimators of the finite population distribution function. *Biometrika* **1992**, *79*, 577–582.
29. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* **1994**, *89*, 846–866.
30. Rao, J.; Kovar, J.; Mantel, H. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **1990**, pp. 365–375.
31. Woodruff, R.S. Confidence intervals for medians and other position measures. *Journal of the American Statistical Association* **1952**, *47*, 635–646.
32. Lohr, S.L. *Sampling: design and analysis*; Chapman and Hall/CRC, 2021.
33. Sitter, R.R.; Wu, C. A note on Woodruff confidence intervals for quantiles. *Statistics & probability letters* **2001**, *52*, 353–358.