mathematics

MDPI

Article

Methods to Counter Self-Selection Bias in Estimations of the Distribution Function and Quantiles

María del Mar Rueda [1,*], Sergio Martínez-Puertas [2] and Luis Castro-Martín [3]

[1] Department of Statistics and O.R. and Institute of Mathematics, University of Granada, 18071 Granada, Spain
[2] Department of Mathematics, University of Almería, 04120 Almería, Spain
[3] Andalusian School of Public Health, University of Granada, 18011 Granada, Spain
* Correspondence: mrueda@ugr.es

Abstract: Many surveys are performed using non-probability methods such as web surveys, social networks surveys, or opt-in panels. The estimates made from these data sources are usually biased and must be adjusted to make them representative of the target population. Techniques to mitigate this selection bias in non-probability samples often involve calibration, propensity score adjustment, or statistical matching. In this article, we consider the problem of estimating the finite population distribution function in the context of non-probability surveys and show how some methodologies formulated for linear parameters can be adapted to this functional parameter, both theoretically and empirically, thus enhancing the accuracy and efficiency of the estimates made.

Keywords: nonprobability surveys; propensity score adjustment; survey sampling; poverty measures

MSC: 62D05

Citation: Rueda, M.d.M.; Martínez-Puertas, S.; Castro-Martín, L. Methods to Counter Self-Selection Bias in Estimations of the Distribution Function and Quantiles. *Mathematics* **2022**, *10*, 4726. https://doi.org/10.3390/math10244726

Academic Editor: Danilo Costarelli

Received: 7 November 2022
Accepted: 8 December 2022
Published: 12 December 2022

1. Introduction

Distribution function estimation is an important topic in survey research. This approach offers valuable benefits in the context of probability surveys and has been the focus of much research attention in recent years. It is especially useful when the underlying goal is to determine the proportion of values of a study variable that are less than or equal to a certain value. For example, knowledge of the distribution function makes it possible to obtain the reliability function, which is commonly used in life data analysis and reliability engineering [1]. Furthermore, the distribution function allows us to examine whether two samples originate from the same population [2].

Additionally, the finite population distribution function can be used to calculate certain parameters, such as population quantiles. In several areas of study [3–5], quantiles are of special interest. For example, rates of extreme pediatric obesity are defined as the body mass index at or above the 99th percentile [6]. In another area, that of ozone concentrations, the 5th percentile is a measure of the baseline condition, while the 95th reflects peak concentration levels [7]. In economics, some variables, such as income, have skewed distributions and in this case quantiles provide a more suitable measure of location than the mean [8,9]. Also in this field, quantiles allow us to obtain measures such as the poverty line and the poverty gap [10,11], as well as inequality parameters indicators such as the headcount index, which measures the proportion of individuals classified as poor within a given population [12]. Other analyses of inequality, such as those focusing on wages or income distribution, also require measures based on percentile ratios [9]. In these cases, estimating the distribution function is also more useful than calculating totals and means [13]. In view of these considerations, some studies have focused on the auxiliary population information available at the estimation stage to gain more accurate values for the distribution function and quantiles [14–17]. One means of incorporating auxiliary information to develop new estimators of the distribution function is to employ the calibration method, which was originally designed to estimate the total

population [18]. An extensive body of research has been conducted in this area, and various implementations of the calibration approach have been applied in the probability survey context to obtain estimators of the distribution function and the quantiles [19–27]. The use of calibration techniques has also been considered for estimating the distribution function when a probability survey is subject to non-response [28].

As part of the global commitment to fight poverty and social exclusion, many government agencies wish to know the proportion of the population living below the poverty line, in order to monitor the effectiveness of their policies [29]. One way to obtain this information is to conduct probabilistic surveys, based on representative samples of the target population. The aim of survey sampling theory is to maximize the reliability of the estimates thus obtained.

For a sample to be considered probabilistic and therefore valid for drawing inferences regarding the population, it must be selected under the assumption that all the individuals in the target population have a known and non-null probability of inclusion.

In recent years, alternative data sources to probabilistic samples have been considered, such as big data and web surveys. These approaches offer certain advantages over traditional probability sampling: estimates in near real time may be obtained, data access is easier, and data collection costs are lower. In these non-traditional methods, the data generating process is different and the subsequent analysis is based on non-probability samples. Despite the above advantages, this method also presents serious issues, especially the fact that the selection procedure for the units included in the sample is unknown and so the estimates obtained may be biased, since the sample itself does not necessarily provide a valid picture of the entire population. In other words, the sample is potentially exposed to self-selection bias [30,31].

Many studies of survey sampling have been undertaken to reduce selection bias in the methods used to estimate population totals and means, and this research has been reviewed in [32–37], among others. The methods considered include inverse probability weighting [38,39], inverse sampling [34], mass imputation [40], doubly robust methods [31], kernel smoothing methods [41], statistical matching combined with propensity score adjustment [42], and calibration combined with propensity score adjustment [39,43]. However, despite the extensive literature available on using calibration techniques to estimate the distribution function and the population mean under conditions of self-selection bias, little attention has been paid to the development of efficient methods for estimating the population distribution function under these conditions.

To address this research gap, we propose a general framework for drawing statistical inferences for the distribution function with non-probability survey samples when auxiliary information is available. We discuss different methods of adjusting for self-selection bias, depending on the type of information available, applying calibration, propensity score, and statistical matching techniques.

The rest of the paper is organized as follows: in Section 2, we review the estimation of the distribution function from probability and non-probability samples, in order to establish the basic framework and the notation employed. In Section 3, we then propose several estimators for the distribution function, based on calibration, propensity score adjustment (PSA) and statistical matching (SM), taking into consideration that the non-smooth nature of the finite population distribution function produces certain complexities, which are resolved in different ways. The properties of the proposed estimators are described in Section 4, after which we present the results obtained from the simulation studies performed with these estimators. In the final section, we summarize the main conclusions drawn and suggest possible lines of further research in this area.

2. Basic Setup for Estimating the Distribution Function

Let U denote a finite population of size N, $U = \{1, \ldots, i, \ldots, N\}$. Let s_V be a self-selected sample of size n_V, self-selected from U. Let y be the variable of interest in the

survey estimation. We assume that y_k is known for all sample units. Our goal is to estimate the distribution function $F_y(t)$ for the study variable y, which can be defined as follows:

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) \tag{1}$$

where $\Delta(\cdot)$ denotes the Heaviside function, given by:

$$\Delta(t - y_k) = \begin{cases} 1 & \text{if } t \geq y_k \\ 0 & \text{if } t < y_k. \end{cases}$$

In the absence of auxiliary information, the distribution function $F_y(t)$ can be estimated by the naive estimator, defined by

$$\widehat{F}_{YNa}(t) = \frac{1}{n} \sum_{k \in s_V} \Delta(t - y_k). \tag{2}$$

If the convenience sample s_V suffers from selection bias, the above estimator will provide biased results.

Let R be an indicator variable of an element being in s_V, such that

$$R_k = \begin{cases} 1 & k \in s_V \\ 0 & k \notin s_V. \end{cases} \tag{3}$$

If we know $\{R_k : k \in U\}$, the error of $\widehat{F}_{YNa}(t)$ will be:

$$\widehat{F}_{YNa}(t) - F_y(t) = \frac{1}{n} \sum_{k \in U} R_k \Delta(t - y_k) - \frac{1}{N} \sum_{k \in U} \Delta(t - y_k) = \frac{1}{f} Cov(R, \Delta(t - y))$$

with $f = n/N$ and $Cov(R, \Delta(t - y)) = \frac{1}{N} \sum_{k \in U} (R_k - \overline{R}_N)(\Delta(t - y_k) - F_y(t))$, being $\overline{R}_N = \frac{1}{N} \sum_{k \in U} R_k$.

By applying the expectation of the mean difference, we obtain the selection bias of the estimator, as follows:

$$B = E_R(\widehat{F}_{YNa}(t) - F_y(t)) = \frac{1}{f} E_R(Cov(R, \Delta(t - y)))$$

where E_R denotes the expectation with respect to the random mechanism for R_k.

The mean squared error is obtained by:

$$ECM = \frac{1}{f^2} E_R(Cov(R, \Delta(t - y))^2) = \frac{1}{f^2} E_R(Corr(R, \Delta(t - y))^2) Var(R) Var(\Delta(t - y)) =$$

$$= E_R(Corr(R, \Delta(t - y))^2) \times \left(\frac{1}{f} - 1 \right) \times Var(\Delta(t - y))$$

because $Var(R) = \frac{1}{N} \sum_{k \in U} (R_k - \overline{R}_N)^2 = f(1 - f)$.

Therefore, a non-probability sampling design with $E_R(Corr(R, \Delta(t - y))^2) \neq 0$ means that the analysis results are subject to selection bias. This is the main problem addressed in our study.

3. Proposed Estimators

The key to successful weighting to eliminate bias in self-selection surveys lies in the use of appropriate auxiliary information. To address this question, let us consider J auxiliary variables $x_1, \ldots, x_J$ and let $\mathbf{x}_k' = (x_{1k}, \ldots, x_{Jk})$ be the vector of auxiliary variables at unit k.

We distinguish three different cases, called InfoTP, InfoES, and InfoES, depending on the information at hand ([44])

- InfoTP: Only the population vector totals of the auxiliary variables, $\sum_U \mathbf{x}_k = \mathbf{X}$, are known.
- InfoES: Information is available at the level of a probability sample conducted on the same target population as the non-probability survey, with good coverage and high response rates. The vector of auxiliary variables $\mathbf{x}_k$ is known for every unit in this reference sample.
- InfoEP: Information is available at the level of the population U: the vector of auxiliary variables $\mathbf{x}_k$ is known for every $k \in U$.

Below, we consider various adjustment methods, depending on the type of information available.

3.1. InfoTP

The calibration method, originally developed by Deville and Särndall [18] for the estimation of totals, can be adapted to estimate the distribution function. This approach enables us to incorporate the auxiliary information available through the auxiliary vector $\mathbf{x}_k$ in several ways [19,20,24–27].

In the case of InfoTP, the calibration can be performed on the totals: given a pseudo-distance $G(.,.)$, and denoting $w_{vk} = N/n_V$, we seek new calibrated weights w_{kc1} that are the solution to the following minimization problem

$$\min_{w_k} \sum_{k \in s_V} G(w_k, w_{vk}) \tag{4}$$

subject to

$$\sum_{k \in s_V} w_k \mathbf{x}_k = \mathbf{X}. \tag{5}$$

The resulting calibrated estimator of the distribution function is given by:

$$\widehat{F}_{Yc1}(t) = \frac{1}{N} \sum_{k \in s_V} w_{kc1} \Delta(t - y_k). \tag{6}$$

Ref. [18] proposes a family of pseudo-distance $G(.,.)$ with which to develop calibration estimators. One of the distances proposed is the chi-square distance given by

$$\Phi_s = \sum_{k \in s_V} \frac{(w_k - w_{vk})^2}{w_{vk} q_k} \tag{7}$$

where q_k is positive weights that are usually assumed as uniform $1/q_k = 1$ although unequal weights $1/q_k$ are sometimes preferred.

The resulting calibrated weights w_{kc1} with the minimization of (7) subject to the conditions (5) are given by:

$$w_{kc1} = w_{vk} + w_{vk} q_k \gamma \cdot \mathbf{x}_k \tag{8}$$

where

$$\gamma = \left(\mathbf{X} - \sum_{k \in s_V} w_{vk} \mathbf{x}_k \right)^T \left(\sum_{k \in s_V} w_{vk} q_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1}$$

In the estimation of totals and means, previous research has shown that the exclusive use of calibration fails to eliminate self-selection bias if this approach is not combined with other methods, such as propensity score adjustment (PSA) [39,43]. Thus, in terms of bias

reduction, the results of the calibration and PSA combination clearly surpass those obtained with only calibration weighting [43].

In order to incorporate methods such as PSA and to develop new estimators that overcome the problems met with the $\widehat{F}_{Yc1}(t)$ estimator, we consider other scenarios as follows.

3.2. InfoES

Let s_R be a probability sample of size n_R selected from U under a probability sampling design (s_R, p_R) in which $\pi_k = \sum_{s_R \ni k} p_R(s_R) > 0$ is the first-order inclusion probability for individual k. The covariates $\mathbf{x}_k$ are common to both samples, but we only have measurements of the variable of interest y for the individuals in the convenience sample. The original design weight of the individual k in the reference (probability) sample is denoted by $w_{Rk} = 1/\pi_k$.

First, we consider a calibration method for reweighting based on the proposal given in [25], calibrating from the pseudo-variable:

$$g_k = \widehat{\beta}^T \mathbf{x}_k \text{ for } k = 1, 2, ...N \tag{9}$$

$$\widehat{\beta} = \left(\sum_{k \in s_V} \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} \cdot \sum_{k \in s_V} \mathbf{x}_k y_k. \tag{10}$$

The new weights w_{kc2} are obtained by minimizing the chi-square distance (7) subject to the following conditions:

$$\frac{1}{N} \sum_{k \in s_V} w_{kc2} \Delta(t_j - g_k) = \frac{1}{N} \sum_{k \in s_R} w_{Rk} \Delta(t_j - g_k) \quad j = 1, 2, \ldots, P \tag{11}$$

where t_j for $j = 1, ...P$ are points chosen arbitrarily and where we assume that $t_1 < t_2 < ... < t_P$ and q_k are positive constants.

The resulting calibrated estimator of the distribution function is given by:

$$\widehat{F}_{Yc2}(t) = \frac{1}{N} \sum_{k \in s_V} w_{kc2} \Delta(t - y_k). \tag{12}$$

in which the calibrated weights w_{kc2} are given by:

$$w_{kc2} = w_{vk} + w_{vk} q_k \frac{\lambda}{N} \Delta(\mathbf{t}_P - g_k) \tag{13}$$

with

$$\lambda = N^2 \cdot \left(\widehat{F}_{GR}(\mathbf{t}_P) - \frac{1}{N} \sum_{k \in s_V} \Delta(\mathbf{t}_P - g_k) \right)^T \left(\sum_{k \in s_V} w_{vk} q_k \Delta(\mathbf{t}_P - g_k) \Delta(\mathbf{t}_P - g_k)^T \right)^{-1}$$

and

$$\Delta(\mathbf{t}_P - g_k)^T = \left(\Delta(t_1 - g_k), \Delta(t_2 - g_k), \ldots, \Delta(t_P - g_k) \right)$$

$$\left(\widehat{F}_{GR}(\mathbf{t}_P) \right)^T = \left(\frac{1}{N} \sum_{k \in s_R} w_{Rk} \Delta(t_1 - g_k), \frac{1}{N} \sum_{k \in s_R} w_{Rk} \Delta(t_2 - g_k), \ldots, \frac{1}{N} \sum_{k \in s_R} w_{Rk} \Delta(t_P - g_k) \right)$$

The calibrated weights (13) and the weights w_{Rk} for the samples s_V and s_R, respectively, give the same estimates for the distribution function of the pseudo-variable g, when evaluated over the set of points t_j.

In the case of InfoES information, the most popular adjustment method in non-probability settings is propensity score adjustment [38,39,43,45–47]. This method, developed by [48], can be used to estimate the distribution function, as described below.

Under PSA, it is assumed that each element of U has a probability (propensity) of being selected for s_V, which can be formulated as

$$\pi_k^v = Pr(R_k = 1|\mathbf{x}_k, y_k) \tag{14}$$

We assume that the response selection mechanism is ignorable and follows a parametric model:

$$\pi_k^v = Pr(R_k = 1|\mathbf{x}_k) = m(\mathbf{x}_k, \lambda), \tag{15}$$

for a known function $m(\cdot)$ with second continuous derivatives with respect to an unknown parameter λ.

We estimate the propensity scores π_k^v by using data from both the self-selection and the probability samples. The maximum likelihood estimator (MLE) of π_k^v is $\hat{\pi}_k^v = m(\hat{\lambda}, \mathbf{x}_k)$, where $\hat{\lambda}$ corresponds to the value of λ that maximizes the pseudo-log-likelihood function:

$$\tilde{l}(\lambda) = \sum_{s_V} log \frac{m(\lambda, \mathbf{x}_k)}{1 - m(\lambda, \mathbf{x}_k)} + \sum_{s_R} \frac{1}{\pi_k} log(1 - m(\lambda, \mathbf{x}_k)). \tag{16}$$

The resulting propensities can then be used to calculate new weights, $w_k^{PSA} = \frac{1}{\hat{\pi}_k^v}$. Thus, we define the inverse propensity weighting estimator of the distribution function as:

$$\widehat{F}_{YIPS}(t) = \frac{1}{N} \sum_{k \in s_V} w_k^{PSA} \Delta(t - y_k). \tag{17}$$

Another PSA-based estimator can be obtained using the weights $w_k^{PSA2} = \frac{1 - \hat{\pi}_k^v}{\hat{\pi}_k^v}$ [49]. In this respect, Refs. [39,47] proposed other PSA weights whereby the combined sample $(s_V \cup s_R)$ is grouped into g equally-sized strata of similar propensity scores from which an average propensity is calculated for each group.

The estimator (17) can be obtained as a special case of the general framework on inference for the general parameter proposed in [50]. The latter authors present an estimator that uses the propensity score for each individual in the survey weighted by the estimating equation under logistic regression, thus obtaining the asymptotic variance of the estimator.

A third approach to dealing with InfoES information is that of statistical matching, by which imputed values are created for all elements in the probability sample. This method was introduced by [40] and is based on modeling the relationship between y_k and $\mathbf{x}_k$, using the self-selected sample s_V to predict y_k for the probability sample. The question then is how to predict the values y_k.

To do so, let us assume a working population model $E_m(y/\mathbf{x}) = M(\mathbf{x}, \beta)$ where β is the unknown parameter. We further assume that the population model holds for the sample s_V. Using the data from this sample, we can obtain an estimator $\hat{\beta}_v$ which is consistent for β under the model assumed. From $\hat{\beta}_v$, we then propose the matching estimator for the distribution function as:

$$\widehat{F}_{YSM}(t) = \frac{1}{N} \sum_{k \in s_R} \Delta(t - \hat{y}_k)/\pi_k \tag{18}$$

where $\hat{y}_k = M(\mathbf{x}, \hat{\beta}_v)$ is the predicted value of y_k under the above model. The estimator (18) is consistent if the model for the study variable is correctly specified.

A more complex estimator for the distribution function can be constructed using the idea of double robust estimation [31], which is based on the following considerations. Firstly, the propensity score adjusted estimator (17) requires that the propensity score model be correctly specified. Moreover, the imputation-based estimator (18) requires that the working population model be correctly specified. An estimator is called doubly robust if the estimator is consistent whenever one of these two models is correctly specified [51]. Hence, the double robust estimator of the distribution function is defined as:

$$\widehat{F}_{YDR}(t) = \frac{1}{N}\Big(\sum_{k\in s_R} \Delta(t-\hat{y}_k)/\pi_k + \sum_{k\in s_V} w_k^{PSA}(\Delta(t-y_k) - \Delta(t-\hat{y}_k)\Big). \tag{19}$$

The estimator (19) is double robust because it is consistent if either the model for the participation probabilities or the model for the study variable is correctly specified.

3.3. InfoEP

In the case of InfoEP, an initial possibility is to consider a similar calibrated estimator, based on the proposal given in [25]. The new weights w_{kc3} are obtained by minimizing the chi-square distance 7 subject to the following conditions:

$$\frac{1}{N}\sum_{k\in s_V} w_{kc3}\Delta(t_j - g_k) = F_g(t_j) \quad j = 1, 2, \ldots, P \tag{20}$$

where $F_g(t_j)$ is the finite distribution function of g at the points $t_j, \quad j = 1, 2, \ldots, P$.

The resulting calibrated estimator of the distribution function is:

$$\widehat{F}_{Yc3}(t) = \frac{1}{N}\sum_{k\in s_V} w_{kc3}\Delta(t - y_k) \tag{21}$$

where the calibrated weights w_{kc3} are given by:

$$w_{kc3} = w_{vk} + w_{vk}q_k\frac{\theta}{N}\Delta(\mathbf{t}_P - g_k) \tag{22}$$

with

$$\theta = N^2 \cdot \Big(F_g(\mathbf{t}_P) - \frac{1}{N}\sum_{k\in s_V}\Delta(\mathbf{t}_P - g_k)\Big)^T \left(\sum_{k\in s_V} w_{vk}q_k\Delta(\mathbf{t}_P - g_k)\Delta(\mathbf{t}_P - g_k)^T\right)^{-1}$$

$$\Big(F_g(\mathbf{t}_P)\Big)^T = \Big(F_g(t_1), F_g(t_2), \ldots, F_g(t_P)\Big).$$

$\widehat{F}_{Yc3}(t)$ gives perfect estimates for the distribution function of the pseudo-variable g, when evaluated over the set of points $t_j, \quad j = 1, 2, \ldots, P$.

We define a model-based estimator based on the non-probability sample as

$$\widehat{F}_{YDR2}(t) = \frac{1}{N}\Big(\sum_{k\in U - s_V} \Delta(t-\hat{y}_k) + \sum_{k\in s_V}\Delta(t-y_k)\Big) \tag{23}$$

and a model-assisted estimator by

$$\widehat{F}_{YDR3}(t) = \frac{1}{N}\Big(\sum_{k\in U} \Delta(t-\hat{y}_k) + \sum_{k\in s_V} w^{PSA}(\Delta(t-y_k) - \Delta(t-\hat{y}_k)\Big) \tag{24}$$

4. Properties of Proposed Estimators

When estimating the distribution function, the estimator considered $\widehat{F}_Y(t)$ should satisfy the following distribution function properties:

(i) $\widehat{F}_Y(t)$ should be continuous on the right;

(ii) $\widehat{F}_Y(t)$ should be monotonically nondecreasing;

(iii) $\lim_{t\to-\infty} \widehat{F}_Y(t) = 0$;

(iv) $\{\lim_{t\to+\infty} \widehat{F}_Y(t) = 1$.

If an estimator of the distribution function $\widehat{F}_Y(t)$ is a genuine distribution function, i.e., $\widehat{F}_Y(t)$ meets the above conditions, it can be used directly for estimating the quantiles [16]. Specifically, the quantile α can be estimated as:

$$\widehat{Q}_\alpha = inf\{t : \widehat{F}_Y(t) \geq \alpha\} = \widehat{F}_Y^{-1}(\alpha)$$

Since the Heaviside function is continuous on the right, it is clear that all the proposed estimators satisfy conditions (i) and (iii).

In general, however, estimator $\widehat{F}_{yc1}(t)$ does not satisfy conditions (ii) or (iv). In order to meet condition (ii), let us consider the specific pseudo-distances $G(.,.)$ that guarantee positive calibrated weights $w_{kc1} > 0$. In this respect, Ref. [18] proposed some pseudo-distances which always produce positive weights whilst avoiding extremely large ones. Some of these pseudo-distances may be considered in estimator $\widehat{F}_{yc1}(t)$ in order to satisfy condition (ii). In addition, to meet condition (iv), we can add the constraint:

$$N = \sum_{k \in s_V} w_k \tag{25}$$

to condition (5).

Similarly, conditions (ii) and (iv are not generally met by the estimators $\widehat{F}_{Yc2}(t)$ or $\widehat{F}_{Yc3}(t)$). Regarding condition (ii), and following [25], the weights w_{kc2} and w_{kc3} are always positive if $q_k = c$ for all units in the population. Thus, under the usual uniform choice $1/q_k = 1$, both estimators satisfy condition (ii). To meet condition (iv), in the case of estimator $\widehat{F}_{Yc2}(t)$, we can add the constraint (25) to the conditions (11), while, for estimator $\widehat{F}_{Yc3}(t)$, we can take a value t_P that is large enough so $F_g(t_P) = 1$.

The estimator $\widehat{F}_{YIPS}(t)$ based on the weights w_k^{PSA} verifies condition (ii) if weights $w_k^{PSA} \geq 0$, whereas if $\widehat{F}_{YIPS}(t)$ is based on w_k^{PSA2}, then it meets condition (ii) when $w_k^{PSA} \leq 1$. Consequently, if $0 \leq w_k^{PSA} \leq 1$, the estimator $\widehat{F}_{YIPS}(t)$ based on both w_k^{PSA} and w_k^{PSA2} satisfies condition (ii). Thus, through the model selected to estimate propensities $m(\lambda, \mathbf{x}_k)$, condition (ii) can be met. For example, an extended option in the estimation of propensities is that of the logistic regression model

$$m(\lambda, \mathbf{x}_k) = \frac{exp(\lambda^T \mathbf{x}_k)}{1 + exp(\lambda^T \mathbf{x}_{ik})}$$

that verifies the condition $0 \leq w_k^{PSA} \leq 1$. Hence, if we choose this model, condition (ii) is met by $\widehat{F}_{YIPS}(t)$ regardless of whether we use the weight w_k^{PSA} or the weight w_k^{PSA2}.
To ensure that condition (iv) is met with the estimator $\widehat{F}_{YIPS}(t)$, it can be divided by the sum of weights, that is, $\sum_{k \in s_V} w_k^{PSA}$ or $\sum_{k \in s_V} w_k^{PSA2}$.

Estimator $\widehat{F}_{YSM}(t)$ satisfies condition (ii) but not condition (iv). To ensure the latter, again we can divide $\widehat{F}_{YSM}(t)$ by the sum of its weights, that is, $\sum_{s_R} \pi_k$.

Finally, whereas the estimator $\widehat{F}_{YDR2}(t)$ satisfies all the conditions, $\widehat{F}_{YDR}(t)$ and $\widehat{F}_{YDR3}(t)$ do not meet conditions (ii) or (iv). Condition (iv) can be met by both $\widehat{F}_{YDR}(t)$ and $\widehat{F}_{YDR3}(t)$ when they are divided by the sum of their respective weights, but these estimators, in general, are not monotonic non-decreasing functions and therefore are not genuine distribution functions. In both cases, we might consider the general procedure described in [16] to obtain a monotonous non-decreasing version of the estimators $\tilde{F}_{YDR}(t)$ and $\tilde{F}_{YDR3}(t)$. However, this procedure always increases the computational cost when estimating quantiles.

5. Simulation Study

In this section, we conduct a Monte Carlo study to compare the efficiency of the estimators presented in Section 3.2. The simulation study was programmed in R and Python. New code was developed to calculate the estimator considered. Python was only used for training and applying the machine learning models in order to take advantage of the package *Optuna* [52] for hyperparameter optimization. However, R was chosen as the main programming language since the functions *wtd.quantile*, from the package *reldist* [53], and *qgeneric*, from the package *flexsurv* [54], facilitate the implementation of custom quantiles. To show that the superiority of some estimators depends on the data,

we define various setups based on different sampling strategies for the probability and nonprobability samples. In this analysis, only InfoES information is used.

5.1. Data

The dataset used in the simulation was collected between 2011 and 2012, in the Spanish Life Conditions Survey [55]. Using criteria harmonized for all European Union countries, the Living Conditions Survey generates a reference source of statistics on income distribution and social exclusion within Europe. The dataset was filtered to rule out individuals and variables with large quantities of missing data. Following this procedure, the resulting pseudopopulation had a size of $N = 28210$.

The following variables were used in the simulation:

- Demographics

- COM: 1 if the individual has a computer at home, and 0 otherwise;
- SEX: 1 if the individual is male, and 0 otherwise;
- AGE: the individual's age in years;
- $AREAME$: 1 if the individual lives in a medium-density population area, and 0 otherwise;
- $AREALOW$: 1 if the individual lives in a low-density population area, and 0 otherwise.

- Analysis variable

- INC: Household expenses in EUR.

Let us consider two setups. In the first, the sampling procedure is the same as that used to select the sample in the Spanish Life Conditions Survey: the probability sample is obtained by stratified cluster sampling, whereby the strata are defined by the NUTS2 regions and the clusters are composed of the households within these regions, extracted with probabilities proportional to the household size. The number of households to be selected, m, is estimated by dividing n_R (the sample size of s_{R1}) by the mean household size. For $n_R = 2000$, $m = 902$. According to this procedure, the final sample size of s_{R1} is $n_{R1} = 2003$.

In the second setup, the reference probability sample is drawn by Midzuno sampling with probabilities proportional to the minimum household income necessary for basic subsistence.

To generate the nonprobability sample, s_V, the following scenarios were considered:

1. SC1: Simple random sampling from the population with $COM = 1$
2. SC2: Unequal probability sampling from the full pseudopopulation, where the probability of selection for the i-th individual, p_i, is given as follows:

$$p_i = \frac{1}{1 + exp(-2COM + 0.2SEX + 0.01AGE + 0.2AREAME + 0.4AREALOW)} \tag{26}$$

3. SC3: Unequal probability sampling from the full pseudopopulation, where the probability of selection for the i-th individual, p_i, is given as follows:

$$p_i = (AGE - 1925)^3 / (1995 - 1925)^3 \tag{27}$$

These participation mechanisms create weights with different models and levels of variability. Figures 1–3 show the resulting histogram of propensities.

By this procedure, we obtained nonprobability samples with sizes n_V = 2000, 4000 and 6000.

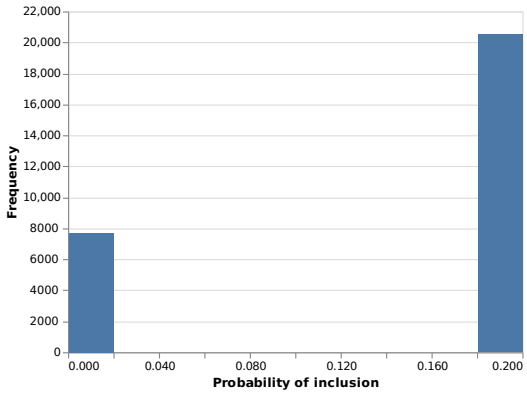

Figure 1. Histogram of population propensities in SC1.

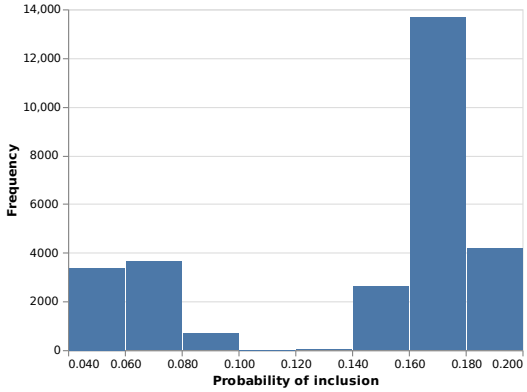

Figure 2. Histogram of population propensities in SC2.

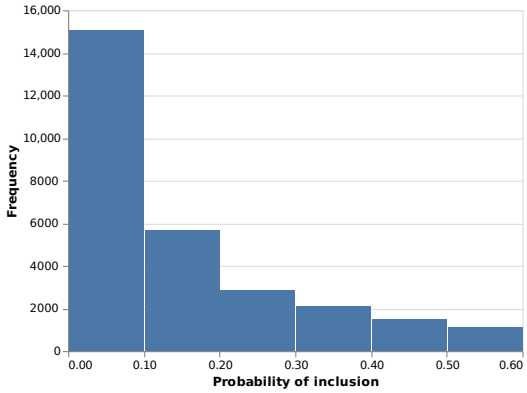

Figure 3. Histogram of population propensities in SC3.

5.2. Simulation

In each simulation, the following parameters were estimated:

- The quantiles $Q_{0.25}$, $Q_{0.5}$ and $Q_{0.75}$.
- The distribution function $F_y(t)$ at points $Q_{0.25}$, $Q_{0.5}$ and $Q_{0.75}$.

The following methods for estimating these parameters were compared:

- Naive estimator, using the sample distribution function of the s_V sample to draw inferences.
- The proposed calibrated estimator $\widehat{F}_{Yc2}(t)$ where t_j for $j = 1, 2, 3$ corresponds to $Q_{0.25}$, $Q_{0.5}$, and $Q_{0.75}$.
- The proposed PSA estimator $\widehat{F}_{YIPS}(t)$.
- The proposed SM estimator $\widehat{F}_{YSM}(t)$.

- The proposed DR estimator $\widehat{F}_{YDR}(t)$.

All five demographic variables were considered potential predictors of propensities, and predicted values y_i for both logistic and linear regression models. In addition, a state-of-the-art machine learning method, XGBoost [56], was used as an alternative to these two models in order to evaluate the effect of the method used to estimate propensities and predict values. Refs. [57,58] show that this technique can improve the representativity of self-selection surveys with respect to other prediction methods.

The quantile α is estimated as follows:

$$\widehat{Q}_\alpha = inf\{t : \widehat{F}_Y(t) \geq \alpha\} = \widehat{F}_Y^{-1}(\alpha)$$

where $\widehat{F}_Y$ is one of the five above estimators of F_y.

One thousand simulations were run for each context. The resulting mean bias, standard deviation, and root mean square error were measured in relative numbers to make them comparable across different scenarios. The formulas used for their calculation were:

$$RBias\ (\%) = \left| \frac{\sum_{i=1}^{1000} \hat{\theta}^{(i)}}{1000} - \theta_N \right| \cdot \frac{100}{\theta_N} \tag{28}$$

$$RStandard\ deviation\ (\%) = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\theta}^{(i)} - \hat{\bar{\theta}})^2}{999}} \cdot \frac{100}{\theta_N} \tag{29}$$

$$RMSE\ (\%) = \sqrt{RBias^2 + RSD^2} \tag{30}$$

where $\hat{\theta}^{(i)}$ is the estimation of a parameter θ_N in the i-th simulation and $\hat{\bar{\theta}}$ is the mean of the 1000 estimations.

5.3. Results

The relative bias of estimators is shown in Table 1, for all scenarios and sample sizes.

Table 1. Bias (%) for each reference probability sample, parameter, non-probability sampling and size, method and machine learning model (linear/logistic regression or XGBoost).

			Stratified						**Proportional**					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC1 2000	Naive		23.3	17.9	13.4	−32.6	−20.9	−10.0	23.3	17.9	13.4	−32.6	−20.9	−10.0
	Cal	Reg	6.7	12.9	17.7	−0.6	−17.2	−20.7	−10.3	2.0	12.0	40.7	1.4	−5.9
		XGB	0.2	3.6	9.8	7.1	4.2	−0.3	1.1	6.3	9.8	5.5	1.0	−0.4
	PSA	Reg	16.8	14.6	11.1	−23.3	−16.5	−8.2	21.5	18.0	13.9	−30.0	−20.4	−10.1
		XGB	19.6	13.1	10.5	−28.1	−15.6	−8.8	28.4	20.9	12.5	−38.2	−24.4	−10.5
	SM	Reg	-4×10^8	25.8	3×10^8	−0.6	−17.3	−20.7	-2×10^9	−0.8	4×10^7	40.7	1.4	−5.9
		XGB	−4.1	−4.2	0.6	7.3	4.1	−0.4	−2.3	−1.2	0.6	5.7	0.9	−0.4
	DR	Reg	-4×10^8	25.8	3×10^8	−0.6	−17.2	−20.7	-2×10^9	−0.9	4×10^7	40.7	1.4	−5.9
		XGB	−4.1	−4.3	0.5	7.3	4.2	−0.3	−2.4	−1.2	0.6	5.6	1.0	−0.4
SC1 4000	Naive		23.3	18.1	13.6	−32.7	−21.0	−10.1	23.3	18.1	13.6	−32.7	−21.0	−10.1
	Cal	Reg	−6.3	2.9	13.7	30.6	−1.6	−10.2	−11.2	1.4	11.7	42.9	2.6	−5.2
		XGB	0.2	3.7	9.9	6.9	4.2	−0.4	1.2	6.3	9.9	5.2	1.0	−0.5
	PSA	Reg	16.8	14.6	11.3	−23.4	−16.6	−8.3	21.5	18.0	13.9	−30.0	−20.4	−10.1
		XGB	19.9	14.0	13.0	−28.0	−16.4	−10.4	28.1	22.1	13.7	−37.1	−25.0	−11.4
	SM	Reg	-2×10^9	4.3	2×10^8	30.6	−1.6	−10.2	-2×10^9	−2.3	6×10^7	42.9	2.5	−5.2
		XGB	−4.2	−4.2	0.6	7.2	4.1	−0.4	−2.2	−1.2	0.6	5.5	0.9	−0.5
	DR	Reg	-2×10^9	4.3	2×10^8	30.6	−1.6	−10.2	-2×10^9	−2.4	6×10^7	42.9	2.6	−5.2
		XGB	−4.1	−4.2	0.6	7.1	4.2	−0.3	−2.3	−1.2	0.6	5.4	1.0	−0.5

Table 1. *Cont.*

			Stratified						Proportional					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC1 6000	Naive		23.3	18.0	13.5	−32.7	−20.9	−10.0	23.3	18.0	13.5	−32.7	−20.9	−10.0
	Cal	Reg	0.7	8.1	15.8	13.9	−9.9	−15.8	−10.9	1.6	11.7	42.0	2.1	−5.5
		XGB	0.2	3.6	9.7	6.8	4.1	−0.4	1.2	6.2	9.7	5.1	1.0	−0.5
	PSA	Reg	16.8	14.6	11.4	−23.5	−16.5	−8.3	21.4	17.9	13.9	−30.0	−20.3	−10.1
		XGB	20.1	14.6	14.8	−27.6	−17.2	−11.4	27.6	22.5	14.3	−36.0	−25.1	−11.9
	SM	Reg	-7×10^8	15.8	1×10^8	13.9	−10.0	−15.8	-3×10^9	−1.7	9×10^7	42.0	2.1	−5.5
		XGB	−4.3	−4.1	0.6	7.0	4.2	−0.4	−2.1	−1.2	0.6	5.4	1.0	−0.5
	DR	Reg	-7×10^8	15.8	1×10^8	13.9	−9.9	−15.8	-3×10^9	−1.8	9×10^7	42.0	2.1	−5.5
		XGB	−4.2	−4.2	0.6	7.0	4.2	−0.4	−2.3	−1.3	0.6	5.3	1.0	−0.5
SC2 2000	Naive		12.1	10.2	7.8	−18.2	−11.6	−5.6	12.1	10.2	7.8	−18.2	−11.6	−5.6
	Cal	Reg	−1.4	0.1	5.9	6.7	4.0	−0.4	−0.7	3.1	5.9	5.0	1.1	−0.6
		XGB	−1.5	0.0	5.8	7.1	4.2	−0.3	−0.8	3.2	5.9	5.5	1.0	−0.4
	PSA	Reg	−5.2	−4.3	−2.9	9.5	4.9	2.0	−0.2	0.1	−0.2	0.2	0.1	0.2
		XGB	5.4	2.4	2.5	−8.5	−2.4	−2.1	15.2	10.8	5.9	−22.5	−12.3	−4.7
	SM	Reg	−4.3	−4.0	0.6	6.7	4.0	−0.4	−2.1	−1.1	0.6	5.0	1.0	−0.6
		XGB	−4.1	−4.2	0.5	7.2	4.2	−0.3	−2.3	−1.2	0.6	5.6	0.9	−0.4
	DR	Reg	−4.3	−4.0	0.6	6.7	4.0	−0.4	−2.2	−1.2	0.6	5.0	1.1	−0.6
		XGB	−4.1	−4.3	0.5	7.1	4.2	−0.3	−2.4	−1.2	0.6	5.5	1.0	−0.4
SC2 4000	Naive		11.8	10.1	7.7	−17.8	−11.5	−5.6	11.8	10.1	7.7	−17.8	−11.5	−5.6
	Cal	Reg	−1.4	−0.1	5.8	6.6	4.1	−0.4	−0.7	3.0	5.8	5.0	1.1	−0.5
		XGB	−1.6	−0.2	5.7	7.0	4.2	−0.4	−0.9	3.1	5.8	5.4	1.0	−0.4
	PSA	Reg	−5.4	−4.3	−2.9	9.8	4.9	2.0	−0.5	−0.1	−0.4	0.8	0.2	0.3
		XGB	3.5	2.0	3.2	−5.6	−1.8	−2.7	13.9	10.8	6.4	−20.3	−12.1	−5.2
	SM	Reg	−4.3	−4.0	0.6	6.6	4.0	−0.4	−2.1	−1.1	0.6	4.9	1.1	−0.5
		XGB	−4.1	−4.1	0.6	7.3	4.1	−0.4	−2.2	−1.2	0.6	5.6	0.9	−0.5
	DR	Reg	−4.3	−4.0	0.5	6.6	4.1	−0.4	−2.1	−1.2	0.6	5.0	1.1	−0.5
		XGB	−4.1	−4.2	0.6	7.0	4.2	−0.3	−2.4	−1.2	0.6	5.4	1.0	−0.5
SC2 6000	Naive		11.4	9.9	7.4	−17.3	−11.1	−5.4	11.4	9.9	7.4	−17.3	−11.1	−5.4
	Cal	Reg	−1.5	−0.2	5.6	6.6	4.1	−0.4	−0.7	2.9	5.7	4.9	1.1	−0.5
		XGB	−1.6	−0.3	5.6	7.0	4.2	−0.4	−0.9	3.0	5.6	5.3	1.0	−0.5
	PSA	Reg	−5.3	−4.4	−2.8	9.8	4.9	1.9	−0.6	−0.2	−0.5	0.9	0.3	0.4
		XGB	1.8	1.4	3.6	−3.0	−1.2	−3.1	12.3	10.3	6.5	−18.1	−11.3	−5.1
	SM	Reg	−4.3	−4.0	0.6	6.5	4.1	−0.4	−2.1	−1.1	0.6	4.9	1.1	−0.5
		XGB	−4.2	−4.2	0.6	7.3	4.1	−0.4	−2.2	−1.1	0.6	5.6	1.0	−0.5
	DR	Reg	−4.3	−4.0	0.5	6.6	4.1	−0.4	−2.1	−1.2	0.5	4.9	1.1	−0.5
		XGB	−4.1	−4.2	0.6	7.0	4.2	−0.3	−2.3	−1.2	0.6	5.3	1.1	−0.5
SC3 2000	Naive		9.8	8.9	6.9	−14.2	−10.2	−4.9	9.8	8.9	6.9	−14.2	−10.2	−4.9
	Cal	Reg	−2.1	−0.5	5.3	7.4	4.3	−0.3	−1.3	2.6	5.4	5.5	1.1	−0.6
		XGB	−1.9	−0.5	5.3	7.0	4.2	−0.3	−1.2	2.8	5.3	5.5	1.0	−0.4
	PSA	Reg	1.7	0.7	−0.6	−3.0	−0.6	0.5	3.4	2.3	0.7	−6.1	−2.5	−0.4
		XGB	10.5	5.7	5.3	−14.5	−5.6	−4.0	13.1	9.4	6.6	−17.8	−10.4	−5.0
	SM	Reg	−4.5	−4.4	0.6	7.4	4.3	−0.3	−2.5	−1.2	0.6	5.5	1.1	−0.6
		XGB	−4.1	−4.1	0.6	7.1	4.2	−0.3	−2.4	−1.1	0.6	5.6	1.0	−0.4
	DR	Reg	−4.5	−4.5	0.5	7.4	4.3	−0.3	−2.6	−1.2	0.6	5.5	1.1	−0.6
		XGB	−4.1	−4.2	0.5	7.1	4.2	−0.3	−2.5	−1.2	0.6	5.5	1.1	−0.4
SC3 4000	Naive		10.0	9.0	6.8	−14.3	−10.2	−4.9	10.0	9.0	6.8	−14.3	−10.2	−4.9
	Cal	Reg	−1.9	−0.6	5.2	6.8	4.2	−0.4	−1.1	2.6	5.3	5.0	1.1	−0.5
		XGB	−1.8	−0.6	5.2	6.7	4.2	−0.3	−1.2	2.7	5.2	5.3	1.1	−0.5
	PSA	Reg	1.6	1.0	−0.4	−2.7	−1.0	0.3	3.3	2.7	0.7	−5.8	−2.9	−0.5
		XGB	10.2	5.9	5.8	−14.1	−5.6	−4.6	12.7	9.8	7.2	−16.7	−11.0	−5.5
	SM	Reg	−4.3	−4.2	0.5	6.8	4.1	−0.4	−2.1	−1.1	0.6	5.0	1.1	−0.5
		XGB	−4.2	−4.1	0.5	6.9	4.2	−0.4	−2.4	−1.2	0.6	5.4	1.0	−0.5
	DR	Reg	−4.4	−4.2	0.5	6.8	4.2	−0.4	−2.2	−1.2	0.6	5.0	1.1	−0.5
		XGB	−4.1	−4.2	0.6	6.9	4.2	−0.3	−2.5	−1.2	0.6	5.3	1.1	−0.5
SC3 6000	Naive		10.0	9.0	6.7	−14.4	−10.2	−4.9	10.0	9.0	6.7	−14.4	−10.2	−4.9
	Cal	Reg	−1.8	−0.6	5.2	6.6	4.1	−0.4	−1.1	2.6	5.2	4.9	1.1	−0.5
		XGB	−1.8	−0.6	5.2	6.7	4.2	−0.4	−1.2	2.6	5.2	5.2	1.1	−0.5
	PSA	Reg	1.6	0.9	−0.3	−2.9	-0.8	0.2	3.2	2.5	0.8	−5.8	−2.7	−0.5
		XGB	9.8	6.4	6.5	−13.5	−6.0	−5.3	12.6	10.3	7.7	−16.2	−11.4	−6.0
	SM	Reg	−4.3	−4.0	0.5	6.6	4.1	−0.4	−2.1	−1.2	0.6	4.9	1.1	−0.5
		XGB	−4.2	−4.2	0.6	6.9	4.2	−0.4	−2.3	−1.1	0.6	5.4	1.0	−0.5
	DR	Reg	−4.3	−4.0	0.5	6.6	4.1	−0.4	−2.1	−1.2	0.5	4.9	1.1	−0.5
		XGB	−4.2	−4.2	0.6	6.9	4.2	−0.3	−2.4	−1.1	0.6	5.2	1.1	−0.5

These results show that the performance of the methods is very similar for each of the probability sample selections considered. The following comments refer to the first

columns. i.e., those corresponding to the situation in which the probability sample is chosen through a stratified cluster scheme.

The naive estimator for all parameters in Scenario 1, where there is also coverage bias, reflects a very large degree of bias, which is not eliminated by increasing the sample size. The calibration estimator achieves a considerable reduction in the bias when XGBoost is used to predict the values but does not achieve a significant reduction in the bias with linear regression. For some parameters, this bias is even greater than that of the naive estimator.

As expected, in Scenario 1, the PSA-based estimators do not eliminate the self-selection bias, since there is no relationship between the variables of interest and the probability of participation, and the machine learning method used to predict the propensities has little influence. These results are comparable to those reported by [43], who observed that it is important to add covariates related to the study goal in order to make PSA useful.

On the contrary, with the SM method, the ML technique is of determinant importance: the estimators based on linear regression perform very badly, in general, since there is no linear relationship between the values to be predicted and the covariates. However, the XGBoost method works well in the case of nonlinearity and allows us to select the useful covariates in the prediction. A noteworthy finding is the large amount of bias shown by the regression-based estimator for quantiles $Q_{0.25}$ and $Q_{0.75}$, while the version based on XGBoost achieves a very significant error reduction. A very similar pattern of behavior was observed in all cases between the SM and the DR estimators.

With Scenario 2, the estimators present a different behavior pattern. The probability of participation depends on all the covariates, and the PSA method reduces the self-selection bias considerably, in all cases. The ML method has less impact, and the degree of bias reduction achieved is similar in the two methods. Comparable results were obtained with SM and DR, the methods based on calibration. In these cases, the bias reduction in relation to the values obtained with the naive estimator is very large and does not depend on the ML method used. No clear pattern emerged as to which of the methods was the best: for some parameters ($Q_{0.25}$ and $Q_{0.5}$), the calibration method worked better, while for others ($F_y(Q_{0.5})$), the PSA achieved the greatest reduction in bias, and in yet others ($Q_{0.75}$) the best estimates were produced by SM and DR.

In Scenario 3, where the probability of participating depends only on the age covariate, the calibration estimators, DR and SM, also performed well, obtaining a good level of bias reduction. The estimator based on PSA with logistic regression was the best of all in this respect, for all parameters. However, when XGBoost was used, this decrease in bias was not observed in some parameters. This may be due to the fact that this ML method is very sensitive to the choice of hyperparameters and in these simulations the default parameters were chosen and no hyperparameter optimization was performed. Table 2 shows the relative RMSE of these estimators, for each scenario.

Table 2. RMSE (%) for each reference probability sample, parameter, non-probability sampling and size, method, and machine learning model (linear/logistic regression or XGBoost).

			Stratified						**Proportional**					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC1 2000	Naive		23.4	18.0	13.5	32.8	21.0	10.0	23.4	18.0	13.5	32.8	21.0	10.0
	Cal	Reg	24.0	21.8	19.1	54.4	32.2	27.5	21.7	13.4	13.3	59.3	21.6	15.6
		XGB	1.1	3.8	9.9	7.3	4.2	0.4	1.5	6.5	9.9	5.6	1.0	0.5
	PSA	Reg	16.9	14.7	11.2	23.5	16.6	8.3	21.6	18.0	13.9	30.1	20.5	10.1
		XGB	20.3	13.7	11.4	29.0	16.2	9.3	28.7	21.2	12.8	38.7	24.7	10.8
	SM	Reg	9×10^8	45.5	7×10^8	54.4	32.2	27.5	3×10^9	28.4	2×10^8	59.3	21.6	15.6
		XGB	4.2	4.3	0.7	7.5	4.2	0.5	2.4	1.3	0.8	5.8	1.0	0.6
	DR	Reg	9×10^8	45.4	7×10^8	54.4	32.2	27.5	3×10^9	28.5	2×10^8	59.3	21.6	15.6
		XGB	4.2	4.3	0.7	7.4	4.2	0.4	2.5	1.3	0.7	5.7	1.0	0.6

Table 2. *Cont.*

			Stratified						Proportional					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC1 4000	Naive		23.4	18.2	13.6	32.8	21.1	10.1	23.4	18.2	13.6	32.8	21.1	10.1
	Cal	Reg	22.3	16.7	15.3	59.4	25.5	19.8	21.4	12.7	13.0	59.6	20.9	14.8
		XGB	0.8	3.7	9.9	7.0	4.2	0.4	1.4	6.4	10.0	5.3	1.0	0.5
	PSA	Reg	16.9	14.7	11.4	23.5	16.6	8.3	21.5	18.0	13.9	30.1	20.4	10.1
		XGB	20.2	14.3	13.5	28.5	16.8	10.6	28.3	22.2	13.8	37.4	25.1	11.5
	SM	Reg	3×10^9	35.3	5×10^8	59.4	25.5	19.8	3×10^9	27.4	3×10^8	59.6	20.9	14.8
		XGB	4.3	4.2	0.7	7.3	4.2	0.5	2.3	1.2	0.7	5.7	1.0	0.6
	DR	Reg	3×10^9	35.2	5×10^8	59.4	25.5	19.8	3×10^9	27.4	3×10^8	59.6	20.9	14.8
		XGB	4.2	4.2	0.7	7.2	4.2	0.4	2.4	1.3	0.6	5.4	1.0	0.5
SC1 6000	Naive		23.3	18.1	13.5	32.8	20.9	10.0	23.3	18.1	13.5	32.8	20.9	10.0
	Cal	Reg	23.1	19.5	17.4	56.8	29.3	24.2	21.5	13.0	13.0	59.5	21.2	15.1
		XGB	0.6	3.7	9.7	6.8	4.2	0.4	1.3	6.2	9.8	5.2	1.0	0.5
	PSA	Reg	16.9	14.6	11.5	23.5	16.5	8.3	21.4	18.0	13.9	30.1	20.4	10.1
		XGB	20.4	14.8	15.1	27.9	17.4	11.6	27.8	22.6	14.4	36.2	25.2	11.9
	SM	Reg	1×10^9	41.0	3×10^8	56.8	29.3	24.2	4×10^9	27.8	3×10^8	59.5	21.2	15.1
		XGB	4.4	4.2	0.7	7.2	4.2	0.5	2.2	1.2	0.7	5.5	1.0	0.6
	DR	Reg	1×10^9	41.0	3×10^8	56.8	29.3	24.2	4×10^9	27.9	3×10^8	59.5	21.2	15.1
		XGB	4.2	4.2	0.6	7.1	4.2	0.4	2.4	1.3	0.6	5.3	1.1	0.5
SC2 2000	Naive		12.4	10.4	8.0	18.5	11.8	5.8	12.4	10.4	8.0	18.5	11.8	5.8
	Cal	Reg	1.5	1.0	6.0	6.7	4.0	0.4	0.9	3.2	6.1	5.0	1.1	0.6
		XGB	1.7	1.1	6.0	7.2	4.2	0.4	1.1	3.4	6.0	5.6	1.0	0.5
	PSA	Reg	5.6	4.6	3.2	10.2	5.2	2.2	1.8	1.6	1.3	3.2	1.7	0.8
		XGB	7.3	4.8	4.2	11.5	5.2	3.5	15.8	11.3	6.6	23.6	12.9	5.3
	SM	Reg	4.3	4.0	0.7	6.7	4.0	0.4	2.1	1.1	0.6	5.0	1.1	0.6
		XGB	4.2	4.2	0.7	7.4	4.2	0.5	2.4	1.3	0.7	5.8	1.0	0.5
	DR	Reg	4.3	4.0	0.6	6.7	4.0	0.4	2.2	1.2	0.6	5.0	1.1	0.6
		XGB	4.1	4.3	0.7	7.2	4.2	0.4	2.5	1.3	0.7	5.6	1.0	0.5
SC2 4000	Naive		11.9	10.2	7.8	18.0	11.6	5.6	11.9	10.2	7.8	18.0	11.6	5.6
	Cal	Reg	1.5	0.7	5.8	6.6	4.1	0.4	0.8	3.1	5.9	5.0	1.1	0.5
		XGB	1.6	0.8	5.8	7.1	4.2	0.4	1.0	3.2	5.8	5.4	1.1	0.5
	PSA	Reg	5.5	4.5	3.0	10.1	5.0	2.0	1.2	1.1	0.9	2.3	1.2	0.6
		XGB	5.0	3.9	4.1	8.0	3.9	3.4	14.3	11.0	6.8	21.0	12.4	5.4
	SM	Reg	4.3	4.0	0.6	6.6	4.0	0.4	2.1	1.1	0.6	4.9	1.1	0.5
		XGB	4.2	4.2	0.7	7.4	4.1	0.5	2.3	1.3	0.7	5.7	1.0	0.5
	DR	Reg	4.3	4.0	0.5	6.6	4.1	0.4	2.1	1.2	0.6	5.0	1.1	0.5
		XGB	4.1	4.2	0.6	7.1	4.2	0.4	2.4	1.3	0.6	5.5	1.1	0.5
SC2 6000	Naive		11.5	9.9	7.5	17.4	11.2	5.4	11.5	9.9	7.5	17.4	11.2	5.4
	Cal	Reg	1.5	0.6	5.7	6.6	4.1	0.4	0.8	2.9	5.7	4.9	1.1	0.5
		XGB	1.7	0.6	5.6	7.1	4.2	0.4	1.0	3.0	5.7	5.3	1.1	0.5
	PSA	Reg	5.4	4.4	2.9	10.0	5.0	2.0	1.1	0.9	0.8	2.1	1.0	0.6
		XGB	3.3	3.1	4.1	5.4	2.9	3.5	12.7	10.5	6.7	18.6	11.6	5.3
	SM	Reg	4.3	4.0	0.6	6.6	4.1	0.4	2.1	1.1	0.6	4.9	1.1	0.5
		XGB	4.3	4.2	0.6	7.4	4.1	0.4	2.2	1.1	0.6	5.7	1.0	0.5
	DR	Reg	4.3	4.0	0.5	6.6	4.1	0.4	2.1	1.2	0.5	4.9	1.1	0.5
		XGB	4.2	4.2	0.6	7.0	4.2	0.4	2.3	1.2	0.6	5.3	1.1	0.5
SC3 2000	Naive		10.1	9.1	7.1	14.5	10.4	5.1	10.1	9.1	7.1	14.5	10.4	5.1
	Cal	Reg	2.2	1.1	5.4	7.4	4.3	0.4	1.5	2.8	5.5	5.6	1.1	0.6
		XGB	2.1	1.1	5.4	7.1	4.3	0.4	1.4	2.9	5.5	5.6	1.1	0.5
	PSA	Reg	6.1	4.3	3.3	10.3	5.0	2.3	6.1	4.4	3.2	10.5	5.0	2.2
		XGB	11.7	7.1	6.5	16.3	7.2	4.9	14.2	10.1	7.4	19.6	11.3	5.6
	SM	Reg	4.5	4.5	0.6	7.4	4.3	0.4	2.5	1.2	0.7	5.6	1.1	0.6
		XGB	4.2	4.2	0.8	7.3	4.2	0.5	2.4	1.2	0.8	5.7	1.0	0.6
	DR	Reg	4.5	4.5	0.6	7.4	4.3	0.4	2.6	1.2	0.6	5.6	1.1	0.6
		XGB	4.2	4.2	0.7	7.2	4.2	0.4	2.5	1.2	0.7	5.6	1.1	0.5
SC3 4000	Naive		10.1	9.0	6.9	14.4	10.3	4.9	10.1	9.0	6.9	14.4	10.3	4.9
	Cal	Reg	1.9	0.8	5.3	6.8	4.2	0.4	1.2	2.7	5.3	5.0	1.1	0.5
		XGB	1.9	0.9	5.3	6.8	4.2	0.4	1.3	2.7	5.3	5.3	1.1	0.5
	PSA	Reg	3.9	3.0	2.2	6.8	3.3	1.5	4.6	3.6	2.1	7.9	4.0	1.5
		XGB	10.7	6.7	6.4	15.0	6.5	5.1	13.3	10.2	7.5	17.6	11.4	5.8
	SM	Reg	4.3	4.2	0.6	6.8	4.1	0.4	2.2	1.1	0.6	5.0	1.1	0.5
		XGB	4.2	4.2	0.7	7.1	4.2	0.4	2.4	1.2	0.7	5.5	1.0	0.6
	DR	Reg	4.4	4.2	0.5	6.8	4.2	0.4	2.3	1.2	0.6	5.0	1.1	0.5
		XGB	4.2	4.2	0.6	7.0	4.2	0.4	2.5	1.2	0.6	5.4	1.1	0.5

Table 2. *Cont.*

			Stratified						Proportional					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC3 6000	Naive		10.0	9.0	6.8	14.5	10.2	4.9	10.0	9.0	6.8	14.5	10.2	4.9
	Cal	Reg	1.8	0.7	5.2	6.6	4.1	0.4	1.1	2.6	5.3	4.9	1.1	0.5
		XGB	1.9	0.8	5.2	6.8	4.2	0.4	1.3	2.7	5.2	5.2	1.1	0.5
	PSA	Reg	3.1	2.4	1.7	5.5	2.6	1.2	4.0	3.1	1.7	7.0	3.5	1.2
		XGB	10.2	6.9	6.9	14.2	6.6	5.5	12.9	10.5	7.9	16.7	11.6	6.2
	SM	Reg	4.3	4.1	0.5	6.6	4.1	0.4	2.1	1.2	0.6	4.9	1.1	0.5
		XGB	4.2	4.2	0.7	7.0	4.2	0.4	2.4	1.1	0.7	5.4	1.1	0.5
	DR	Reg	4.3	4.1	0.5	6.6	4.1	0.4	2.1	1.2	0.5	4.9	1.1	0.5
		XGB	4.2	4.2	0.6	7.0	4.2	0.4	2.5	1.2	0.6	5.3	1.1	0.5

In Scenario 1, the estimators that use linear or logistic regression are the least efficient, due to the bias that is present. Calibrated SM and DR estimators based on XGBoost improve efficiency by reducing bias. Moreover, the RMSE reduction is very strong in some parameters ($Q_{0.75}$ and $F_y(Q_{0.75})$). However, the PSA-based estimates do not produce a significant reduction in RMSE because the propensities cannot be modeled from the covariates.

In Scenarios 2 and 3, all the proposed methods effectively reduce the error in the estimates, with the exception of PSA with XGBoost in some cases, as discussed above.

To determine whether this problem encountered with the XGBoost method in some situations can be resolved with an appropriate choice of hyperparameters, we repeated the simulation using a hyperparameter optimization process based on the Tree-structured Parzen Estimator (TPE) algorithm [59]. In this procedure, the error is estimated by cross-validation on the logistic loss obtained by each possible model over the training data. Accordingly, this process could be replicated in a real-world scenario.

Tables 3 and 4 show the bias and RMSE values for the estimators with this new simulation for Scenario 2 and Setup 2 (the worst scenario for PSA with the default XGBoost method). Similar results were obtained for all other situations, but for reasons of space they are not shown in this paper.

Table 3. Bias (%) including hyperparameter optimization when overfitting.

			Proportional					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC2 2000	Naive		12.1	10.2	7.8	−18.2	−11.6	−5.6
	Cal	Reg	−0.7	3.1	5.9	5.0	1.1	−0.6
		XGB	−0.8	3.2	5.9	5.5	1.0	−0.4
	PSA	Reg	−0.2	0.1	−0.2	0.2	0.1	0.2
		XGB	15.2	10.8	5.9	−22.5	−12.3	−4.7
		XGB (opt)	2.4	2.7	2.1	−4.1	−2.8	−1.4
	SM	Reg	−2.1	−1.1	0.6	5.0	1.0	−0.6
		XGB	−2.3	−1.2	0.6	5.6	0.9	−0.4
	DR	Reg	−2.2	−1.2	0.6	5.0	1.1	−0.6
		XGB	−2.4	−1.2	0.6	5.5	1.0	−0.4

Table 4. RMSE (%) including hyperparameter optimization when overfitting.

			Proportional					
			$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$	$F_y(Q_{0.25})$	$F_y(Q_{0.5})$	$F_y(Q_{0.75})$
SC2 2000	Naive		12.4	10.4	8.0	18.5	11.8	5.8
	Cal	Reg	0.9	3.2	6.1	5.0	1.1	0.6
		XGB	1.1	3.4	6.0	5.6	1.0	0.5
	PSA	Reg	1.8	1.6	1.3	3.2	1.7	0.8
		XGB	15.8	11.3	6.6	23.6	12.9	5.3
		XGB (opt)	3.0	3.2	2.6	5.2	3.4	1.8
	SM	Reg	2.1	1.1	0.6	5.0	1.1	0.6
		XGB	2.4	1.3	0.7	5.8	1.0	0.5
	DR	Reg	2.2	1.2	0.6	5.0	1.1	0.6
		XGB	2.5	1.3	0.7	5.6	1.0	0.5

These results clearly show that, by optimizing the hyperparameters, we have considerably reduced the bias and error of the estimators.

6. Discussion

In recent years, the use of survey-based online research has expanded considerably. Web surveys are an attractive option in many fields of sociological investigation due to their low fieldwork costs and rapid data collection. However, this survey mode is also subject to many limitations in terms of accurately representing the target population, and the estimates thus obtained are highly likely to present coverage and/or self-selection bias. Various correction techniques, such as calibration, propensity score adjustment, and statistical matching, have been proposed as a means of reducing or eliminating these forms of bias.

Our paper focuses on the question of estimating the distribution function. This issue is important: the distribution function is a basic statistic underlying many others; for purposes such as assessing and comparing finite populations, it can be more revealing than the use of simple means and totals. Indeed, many previous studies have been undertaken to consider how calibration techniques may be applied to the estimation of the distribution function in the context of a probability survey [19–27] and even to overcome the problem of non-response [28]. However, to our knowledge, very few, if any, studies have addressed this issue from the standpoint of a non-probability survey. Accordingly, we analyze the efficiency obtained by certain bias-correction techniques such as calibration, propensity score adjustment, and statistical matching in various situations within a non-probability survey context. In this analysis, we consider the performance of several estimators in terms of reducing self-selection bias, using a representative survey sample as a proxy for the target population. Among the results obtained by the estimators proposed for the distribution function, $\widehat{F}_{yc1}(t)$ needs specific pseudo-distances $G(.,.)$ in order to satisfy condition (ii). The estimator $\widehat{F}_{YDR2}(t)$ is always a genuine distribution function and under favorable conditions, the estimators $\widehat{F}_{Yc2}(t)$ and $\widehat{F}_{Yc3}(t)$ also obtain a genuine distribution function. Moreover, with minor modifications, the estimators $\widehat{F}_{YIPS}(t)$ (under a logistic regression model) and $\widehat{F}_{YSM}(t)$ also satisfy the distribution function conditions. On the other hand, the estimators $\widehat{F}_{YDR}(t)$ and $\widehat{F}_{YDR3}(t)$ are not generally monotonically nondecreasing functions, and therefore when estimating quantiles, an additional process, which increases the computational cost, must be applied. All the estimators included in our proposal can be used under linear and nonlinear models. Self-evidently, $\widehat{F}_{YIPS}(t)$, $\widehat{F}_{YSM}(t)$, $\widehat{F}_{YDR}(t)$, $\widehat{F}_{YDR2}(t)$, and $\widehat{F}_{YDR3}(t)$ are applicable to linear or nonlinear models. While the calibrated estimators $\widehat{F}_{Yc2}(t)$ and $\widehat{F}_{Yc3}(t)$ assume a linear model, due to the pseudo-variable g_k, the combination with XGBoost enables them to be used with other models too. Furthermore, $\widehat{F}_{Yc2}(t)$ and $\widehat{F}_{Yc3}(t)$ can cover the nonlinear case through the procedure described in [60]. The behavior of all these estimators is demonstrated through simulation studies.

Although further investigation is needed, our results show that self-selection bias can be greatly reduced by any of the four methods considered, particularly when appropriate covariates and a valid machine learning technique are used, both in estimating propensities and in predicting values. However, our investigation did not enable us to determine which method is best in all situations. Specifically, for each parameter and bias-reduction method, different behavior patterns were obtained. Nevertheless, in general, the calibration method based on XGBoost is fairly efficient in any situation.

Although the methods proposed are shown to be effective in reducing the MSE of quantile and distribution function estimates in various situations, certain limitations exist and must be acknowledged. For the PS-based method, for example, the amount of bias reduction achieved depends on how well the propensity model predictors predict the outcome. If the propensity model is poorly fitted, the PS estimates may even be more biased than naive estimates. This is also the case with estimators based on SM, which need a good model in order to accurately predict the y-values. In addition, for the distribution function, the issue is even more complex; although there is a good linear relationship between y and the covariates x, this relationship is not necessarily transferred to the jump functions $\Delta(t-y)$. In practice, it is often difficult to decide whether the auxiliary

variables contain all the components needed to characterize the selection mechanism and the superpopulation model. Therefore, when selecting the covariates and the function of the model, it is essential to use flexible ML techniques.

Finally, the present study does not address the question of the estimation of variance. Plug-in estimators can be used to construct variance estimators from the expression of the asymptotic variance, but the issue is not simple, as the variance depends on the probability of the sample s_R being selected and on the selection mechanism described by the propensity model. In estimating the variance for nonlinear parameters, jackknife and bootstrap techniques [61] might be useful and should be considered in future research in this area.

Author Contributions: M.d.M.R., S.M.-P. and L.C.-M. contributed equally to the conceptualization of this study, its methodology, software, and original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministerio de Ciencia, Innovación y Universidades (Grant No. PID2019-106861RB-I00), IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 and FEDER/Junta de Andalucía-Consejería de Transformación Económica, Industria, Conocimiento y Universidades (FQM170-UGR20.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Acal, C.; Ruiz-Castro, J.E.; Aguilera, A.M.; Jiménez-Molinos, F.; Roldán, J.B. Phase-type distributions for studying variability in resistive memories. *J. Comput. Appl. Math.* **2019**, *345*, 23–32. [CrossRef]
2. Alba-Fernández, M.V.; Batsidis, A.; Jiménez-Gamero, M.D.; Jodrá, P. A class of tests for the two-sample problem for count data. *J. Comput. Appl. Math.* **2017**, *318*, 220–229. [CrossRef]
3. Decker, R.A.; Haltiwanger, J.; Jarmin, R.S.; Miranda, J. Declining business dynamism: What we know and the way forward. *Am. Econ. Rev.* **2016**, *106*, 203–207. [CrossRef]
4. Gallagher, C.M.; Meliker, J.R. Blood and urine cadmium, blood pressure, and hypertension: A systematic review and meta-analysis. *Environ. Health Perspect.* **2010**, *118*, 1676–1684. [CrossRef] [PubMed]
5. Medialdea, L.; Bogin, B.; Thiam, M.; Vargas, A.; Marrodán, M.D.; Dossou, N.I. Severe acute malnutrition morphological patterns in children under five. *Sci. Rep.* **2021**, *11*, 4237. [CrossRef] [PubMed]
6. Vander Wal, J.S.; Mitchell, E.R. Psychological complications of pediatric obesity. *Pediatr. Clin.* **2011**, *58*, 1393–1401. [CrossRef]
7. Wilson, R.C.; Fleming, Z.L.; Monks, P.S.; Clain, G.; Henne, S.; Konovalov, I.B.; Menut, L. Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996, Äì2005. *Atmos. Chem. Phys.* **2012**, *12*, 437–454. [CrossRef]
8. Decker, R.; Haltiwanger, J.; Jarmin, R.; Miranda, J. The role of entrepreneurship in US job creation and economic dynamism. *J. Econ. Perspect.* **2014**, *28*, 3–24. [CrossRef]
9. Dickens, R.; Manning, A. Has the national minimum wage reduced UK wage inequality? *J. R. Stat. Soc. Ser. A (Stat. Soc.)* **2004**, *167*, 613–626. [CrossRef]
10. De Haan, J.; Pleninger, R.; Sturm, J.E. Does financial development reduce the poverty gap? *Soc. Indic. Res.* **2022**, *161*, 1–27. [CrossRef]
11. Jolliffe, D.; Prydz, E.B. Estimating international poverty lines from comparable national thresholds. *J. Econ. Inequal.* **2016**, *14*, 185–198. [CrossRef]
12. Martínez, S.; Illescas, M.; Martínez, H.; Arcos, A. Calibration estimator for Head Count Index. *Int. J. Comput. Math.* **2020**, *97*, 51–62. [CrossRef]
13. Sedransk, N.; Sedransk, J. Distinguishing among distributions using data from complex sample designs. *J. Am. Stat. Assoc.* **1979**, *74*, 754–760. [CrossRef]
14. Chambers, R.L.; Dunstan, R. Estimating distribution functions from survey data. *Biometrika* **1986**, *73*, 597–604. [CrossRef]
15. Chen, J.; Wu, C. Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Stat. Sin.* **2002**, *12*, 1223–1239.
16. Rao, J.N.K.; Kovar, J.G.; Mantel, H.J. On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika* **1990**, *77*, 365–375. [CrossRef]
17. Silva, P.L.D.; Skinner, C.J. Estimating distribution functions with auxiliary information using poststratification. *J. Off. Stat.* **1995**, *11*, 277–294.
18. Deville, J.C.; Särndal, C.E. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382. [CrossRef]
19. Arcos, A.; Martínez, S.; Rueda, M.; Martínez, H. Distribution function estimates from dual frame context. *J. Comput. Appl. Math.* **2017**, *318*, 242–252. [CrossRef]
20. Harms, T.; Duchesne, P. On calibration estimation for quantiles. *Surv. Methodol.* **2006**, *32*, 37–52.

21. Martínez, S.; Rueda, M.; Arcos, A.; Martínez, H. Optimum calibration points estimating distribution functions. *J. Comput. Appl. Math.* **2010**, *233*, 2265–2277. [CrossRef]
22. Martínez, S.; Rueda, M.; Martínez, H.; Arcos, A. Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *J. Comput. Appl. Math.* **2017**, *318*, 444–459. [CrossRef]
23. Martínez, S.; Rueda, M.; Illescas, M. The optimization problem of quantile and poverty measures estimation based on calibration. *J. Comput. Appl. Math.* **2022**, *45*, 113054. [CrossRef]
24. Mayor-Gallego, J.A.; Moreno-Rebollo, J.L.; Jiménez-Gamero, M.D. Estimation of the finite population distribution function using a global penalized calibration method. *AStA Adv. Stat. Anal.* **2019**, *103*, 1–35. [CrossRef]
25. Rueda, M.; Martínez, S.; Martínez, H.; Arcos, A. Estimation of the distribution function with calibration methods. *J. Stat. Plan. Inference* **2007**, *137*, 435–448. [CrossRef]
26. Singh, H.P.; Singh, S.; Kozak, M. A family of estimators of finite-population distribution function using auxiliary information. *Acta Appl. Math.* **2008**, *104*, 115–130. [CrossRef]
27. Wu, C. Optimal calibration estimators in survey sampling. *Biometrika* **2003**, *90*, 937–951. [CrossRef]
28. Rueda, M.; Martínez, S.; Illescas, M. Treating nonresponse in the estimation of the distribution function. *Math. Comput. Simul.* **2021**, *186*, 136–144. [CrossRef]
29. Bradshaw, J.; Mayhew, E. Understanding extreme poverty in the European Union. *Eur. J. Homelessness* **2010**, *4*, 171–186.
30. Bethlehem, J. Selection Bias in Web Surveys. *Int. Stat. Rev.* **2010**, *78*, 161–188. [CrossRef]
31. Chen, Y.; Li, P.; Wu, C. Doubly Robust Inference with Nonprobability Survey Samples. *J. Am. Stat. Assoc.* **2019**, *115*, 2011–2021. [CrossRef]
32. Beaumont, J.F. Are probability surveys bound to disappear for the production of official statistics? *Surv. Methodol.* **2020**, *46*, 1–29.
33. Buelens, B.; Burger, J.; van den Brakel, J.A. Comparing Inference Methods for Non-probability Samples. *Int. Stat. Rev.* **2018**, *86*, 322–343. [CrossRef]
34. Kim, J.K.; Wang, Z. Sampling techniques for big data analysis. *Int. Stat. Rev.* **2019**, *87*, S177–S191. [CrossRef]
35. Rao, J.N.K. On Making Valid Inferences by Integrating Data from Surveys and Other Sources. *Sankhya B*, **2020**, *83*, 242–272. [CrossRef]
36. Valliant, R. Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **2020**, *8*, 231–263. [CrossRef]
37. Yang, S.; Kim, J.K. Statistical data integration in survey sampling: A review. *Jpn. J. Stat. Data Sci.* **2020**, *3*, 625–650. [CrossRef]
38. Lee, S. Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J. Off. Stat.* **2006**, *22*, 329–349.
39. Lee, S.; Valliant, R. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociol. Method Res.* **2009**, *37*, 319–343. [CrossRef]
40. Rivers D. Sampling for Web Surveys. In Proceedings of the Joint Statistical Meetings, Salt Lake City, UT, USA, 29 July–2 August 2007.
41. Wang, L.; Graubard, B.I.; Katki, H.A.; Li, Y. Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. *J. R. Stat. Soc. Ser. A* **2020**, *183*, 1293–1311. [CrossRef]
42. Castro-Martín, L.; Rueda, M.D.M.; Ferri-García, R. Combining statistical matching and propensity score adjustment for inference from non-probability surveys. *J. Comput. Appl. Math.* **2021**, *404*, 113414. [CrossRef]
43. Ferri-García, R.; Rueda, M.M. Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT—Stat. Oper. Res. Trans.* **2018**, *42*, 1–10.
44. Rueda, M.; Ferri-García, R.; Castro. L. The R package NonProbEst for estimation in non-probability survey. *R J.* **2020**, *12*, 406–418. [CrossRef]
45. Elliott, M.R., Valliant, R. Inference for Nonprobability Samples. *Stat. Sci.* **2017**, *32*, 249–264. [CrossRef]
46. Ferri-García, R.; Rueda, M.D.M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE* **2020**, *15*, e0231500. [CrossRef]
47. Valliant, R.; Dever, J.A. Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociol. Method Res.* **2011**, *40*, 105–137. [CrossRef]
48. Rosenbaum, P.R.; Rubin, D.B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
49. Schonlau, M.; Couper, M. Options for Conducting Web Surveys. *Stat. Sci.* **2017**, *32*, 279–292. [CrossRef]
50. Castro-Martín, L.; Rueda, M.D.M.; Ferri-García, R. Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment. *Mathematics* **2020**, *8*, 2096. [CrossRef]
51. Wu, C.; Thompson, M.E. *Sampling Theory and Practice*; Springer Nature: Cham, Switzerland, 2020.
52. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
53. Handcoc, M.S. Relative Distribution Methods; Version 1.7-1. 2022. Available online: https://CRAN.R-project.org/package=reldist (accessed on 20 October 2022).
54. Jackson, C.H. flexsurv: A platform for parametric survival modeling in R. *J. Stat. Softw.* **2016**, *70*, i08. [CrossRef]
55. National Institute of Statistics. *Life Conditions Survey—Microdata*; National Institute of Statistics: Washington, DC, USA, 2012.

56. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
57. Castro-Martín, L.; Rueda, M.D.M.; Ferri-García, R. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques. *Mathematics* **2020**, *8*, 879. [CrossRef]
58. Castro-Martín, L.; Rueda, M.D.M.; Ferri-García, R.; Hernando-Tamayo, C. On the Use of Gradient Boosting Methods to Improve the Estimation with Data Obtained with Self-Selection Procedures. *Mathematics* **2021**, *9*, 2991. [CrossRef]
59. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for hyper-parameter optimization. In Proceedings of the Advances in Neural Information Processing Systems, Granada, Spain, 12–15 December 2011; Volume 24.
60. Rueda, M.; Sánchez-Borrego, I.; Arcos, A.; Martínez, S. Model-calibration estimation of the distribution function using nonparametric regression. *Metrika* **2010**, *71*, 33–44. [CrossRef]
61. Wolter, K.M. *Introduction to Variance Estimation*, 2nd ed.; Springer Inc.: New York, NY, USA, 2007