A new estimator for a finite population distribution function in the presence of complete auxiliary information

Leo Pasquazzi and Lucio De Capitani

Abstract In this work we propose a new estimator for the finite population distribution function of a study variable that uses knowledge about an auxiliary variable. The new estimator is based on a nonparametric superpopulation model that allows for nonlinear regression functions and not identically distributed error components. It employs two local linear regressions to estimate first the regression function and then the cdf of the error components. We propose two versions of the new estimator: a model-based version and a model-assisted one. Their performance is compared with that of several well-known estimators in a simulation study under simple random without replacement sampling and under Poisson sampling with nonconstant inclusion probabilities. The simulation results show that both versions of the new estimator perform very steadily in a great variety of populations and that they are particularly efficient in populations that are best fitted by a nonlinear regression function with regression residuals that do not follow a definite pattern.

Keywords: distribution function estimation, auxiliary information, model-based estimator, model-assisted estimator, generalized difference estimator, model-calibrated estimator

Leo Pasquazzi
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milan, e-mail: leo.pasquazzi@unimib.it

Lucio De Capitani
Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milan, e-mail: lucio.decapitani1@unimib.it

1 Introduction

Since Chambers and Dunstan's seminal paper [4], several estimators for the finite population distribution function of a study variable in the presence of complete auxiliary information have been proposed. To help realize which estimator is most efficient in a given situation, they are often classified into different groups. The perhaps most important distinction is whether an estimator accounts for the sample design or not. Estimators of the first kind are called *model-assisted*, those of the second kind are referred to as *model-based*. *Model-assisted* estimators are asymptotically design-unbiased regardless of whether an assumed superpopulation model is true or not. Their design-variance will be considerably smaller than that of estimators that ignore auxiliary information if the finite population of interest follows the superpopulation model. *Model-based* estimators are on the other hand asymptotically model-unbiased under an assumed superpopulation model. They may have considerably lower design-variance than *model-assisted* estimators, but they may be badly design-biased if the finite population does not follow the superpopulation model and/or the sample design is very unbalanced.

The estimator proposed by Chambers and Dunstan [4] (henceforth CD-estimator) belongs to the class of *model-based* estimators. Other notable members of this class are the estimators proposed by Kuo [9], Dorfman and Hall [7] (DH-estimator), and Chambers, Dunstan and Wherly [3] (CDW-estimator). These estimators are based on different superpopulation models. The CD-estimator assumes a linear regression relationship between the study and the auxiliary variable and allows the variance of the error components to depend on the auxiliary variable only in known way. The DH-estimator accommodates also nonlinear regression relationships, while allowing just for identically distributed error components. The Kuo and the CDW-estimators incorporate a wider range of superpopulation models where the distribution of the study variable may depend on the auxiliary variable in almost any arbitrary way. The CDW-estimator, in particular, may be viewed as the sum of the CD-estimator and a model-bias correction term to protect against model-misspecification. It is worth noting that the CD- and the DH-estimator provide estimates that are genuine distribution functions with probability 1. The same holds for the Kuo-estimator, provided that the regression weights used in its definition are nonnegative and that they do sum to 1, while the CDW-estimator may well provide estimates that are not distribution functions.

As for the class of *model-assisted* estimators, its most prominent members are either *generalized difference* or *model-calibrated* estimators (see [1, 15] and [18]). Well-known examples of estimators of the first kind are the Rao, Kovar and Mantel estimator [12] and some of those analyzed, among other estimators, by Dorfman and Hall [7]. A drawback of these estimators is that they may provide estimates that are not genuine distribution functions. As for *model-calibrated* estimators, examples are given by the pseudo empirical maximum likelihood estimators [5, 17, 11] and by the estimators proposed in [13, 14]. It is worth noting that *model-calibrated* estimators may be defined as to ensure estimates that are genuine distribution functions with probability 1.

The estimator we propose in the present work is in origin conceived as a *model-based* estimator. To accommodate sample designs with markedly nonconstant inclusion probabilities we also propose a *model-assisted* version in the form of a *generalized difference* estimator. The new estimator is based on a very general superpopulation model that allows for a nonlinear regression function and not identically distributed error components. The idea underlying its definition is actually quite simple: two separate nonparametric regressions are employed to estimate first the mean regression function, and then the cdfs of the error components. Based on the outcome of these regressions, predictions for the nonsampled units are computed which are finally used to estimate the finite population distribution function of interest. In view of the generality of the underlying superpopulation model we expect the new estimator to be a competitor of the Kuo- and the CDW-estimator.

The rest of this work is organized as follows. In Section 2 we review the definitions of several well-known estimators from the literature. In Section 3 we introduce the new estimator and hint at conditions under which it should be more efficient than some of its competitors. In Section 4 we present the results of a simulation study and compare the performance of the new estimator with that of its competitors recalled in Section 2. Conclusions and final remarks end this work in Section 5.

2 Estimators from literature

Let (y_i, x_i) be the values of a study variable Y and an auxiliary variable X for unit i of a finite population, $i = 1, 2, \ldots, N$. Let $F_N(y) := \sum_{i=1}^{N} I(y_i \leq y)$, where $I(\cdot)$ denotes as usual the indicator function, be the population distribution function of the study variable, and let s be the set of labels i of the units included in a sample drawn from the population. The Horvitz-Thompson estimator for $F_N(y)$ is defined as

$$\widehat{F}_{HT}(y) := \frac{1}{N} \sum_{i \in s} d_i I(y_i \leq y),$$

where $d_i := \pi_i^{-1}$ denotes the inverse sample inclusion probabilities. $\widehat{F}_{HT}(y)$ is obviously a design-unbiased estimator for $F_N(y)$. However, in the presence of complete auxiliary information, i.e. if we know the values of the X variable for all population units, more efficient estimators may be defined.

2.1 Model-based estimators

The first proposal of an estimator for $F_N(y)$ that takes advantage of complete auxiliary information is probably due to Chambers and Dunstan [4]. Their estimator is based on the assumption that

$$y_i := x_i\beta + v(x_i)\varepsilon_i, \qquad i = 1,2,\ldots,N, \tag{1}$$

where β is unknown, $v(\cdot)$ is a known positive function, and the error components ε_i are i.i.d. with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Given this superpopulation model, Chambers and Dunstan estimate $F_N(y)$ by

$$\widehat{F}_{CD}(y) = \frac{1}{N}\left\{\sum_{i\in s} I(y_i \le y) + \sum_{i\notin s}\widehat{G}_{CD,i}\right\},$$

where

$$\widehat{G}_{CD,i} := \frac{1}{n}\sum_{j\in s} I\left(\frac{y_j - x_j\widehat{\beta}}{v(x_j)} \le \frac{y - x_i\widehat{\beta}}{v(x_i)}\right), \quad i \notin s$$

are predictors for the unobserved indicator functions $I(y_i \le y)$. In the definition of the $\widehat{G}_{CD,i}$'s, n denotes the sample size, and

$$\widehat{\beta} := \frac{\sum_{j\in s} v^{-2}(x_j)x_j y_j}{\sum_{j\in s} v^{-2}(x_j)x_j^2}$$

is the weighted least squares estimator for β based on the observed sample.

The main drawbacks of the CD-estimator stem from the fact that it is based on a superpopulation model that allows merely for a linear regression function and that it requires the user to specify the function $v(\cdot)$. A modified version of the CD-estimator that accommodates also nonlinear regression functions has been proposed and analyzed by Dorfman and Hall [7]. The underlying superpopulation model assumes that

$$y_i := \mu(x_i) + \varepsilon_i, \qquad i = 1,2,\ldots,N, \tag{2}$$

where $\mu(\cdot)$ is allowed to be any smooth function, and where the error components ε_i are assumed to be i.i.d.. Dorfman and Hall [7] estimate $\mu(\cdot)$ by

$$\widehat{\mu}(x) := \sum_{j\in s} w_j(x)y_j,$$

using Nadaraya-Watson kernel weights in place of $w_j(\cdot)$. Based on the estimated regression function $\widehat{\mu}(\cdot)$, they employ

$$\widehat{G}_{DH,i} := \frac{1}{n}\sum_{j\in s} I(y_j - \widehat{\mu}(x_j) \le y - \widehat{\mu}(x_i)), \quad i \notin s$$

as predictors for the unobserved indicator functions $I(y_i \le y)$. The *model-based* estimator for $F_N(y)$ based on these predictors will be denoted by $\widehat{F}_{DH}(y)$ hereafter.

Kuo [9] assumes a more general superpopulation model than (1) and (2) where it is merely assumed that

$$P(y_i \le y) := p(x_i), \qquad i = 1,2,\ldots,N, \tag{3}$$

for some smooth function $p(\cdot)$. The estimator for $F_N(y)$ proposed by Kuo [9] uses

$$\widehat{p}_{K,i} := \sum_{j\in s} w_j(x_i) I(y_j \le y), \quad i \notin s \tag{4}$$

as predictors for the unobserved indicator functions $I(y_i \le y)$. In place of the $w_j(\cdot)$'s Kuo [9] suggests to use either uniform kernel weights, or gaussian kernel weights, or nearest k-neighbor weights. The *model-based* estimator based on the predictors (4) will be denoted by $\widehat{F}_K(y)$ below.

Chambers, Dorfman and Wherly [3] use nonparametric regression to offset the model-bias that arises in $\widehat{F}_{CD}(y)$ when the model (1) is wrong. The predictors for the unobserved indicator functions $I(y_i \le y)$ implicit in their estimator for $F_N(y)$ are given by

$$\widehat{p}_{CDW,i} := \widehat{G}_{CD,i} + \sum_{j\in s} w_j(x_i) \left[\widehat{p}_{K,i} - \widehat{G}_{CD,j}\right], \quad i \notin s,$$

and the corresponding *model-based* estimator for $F_N(y)$ will be denoted by $\widehat{F}_{CDW}(y)$ in what follows.

Dorfman and Hall [7] derived asymptotic expansions of the model-bias and the model-variance of $\widehat{F}_{DH}(y)$, $\widehat{F}_K(y)$ and $\widehat{F}_{CDW}(y)$. Considering Nadaraya-Watson regression weights

$$w_j(x) = \frac{K\left(\frac{x-x_j}{\lambda}\right)}{\sum_{k\in s} K\left(\frac{x-x_k}{\lambda}\right)}, \qquad \lambda > 0,$$

with any symmetric density function in place of $K(\cdot)$, they show that under mild conditions the leading term of the model-variance is of order $O(n^{-1})$ for all three estimators, as the leading term of the model-variance of $\widehat{F}_{CD}(y)$ (see [2]). Moreover, they show that the model-bias of $\widehat{F}_{CDW}(y)$ achieves the parametric $O(n^{-1})$ rate, as the model-bias of $\widehat{F}_{CD}(y)$, when model (1) is correct, while it is of order $O(\lambda^2) + o((n\lambda)^{-1})$, as the model-bias of $\widehat{F}_K(y)$, under the more general model (3) when model (1) is wrong. The model-bias of $\widehat{F}_{DH}(y)$ goes to zero slightly slower: it is of order $O(\lambda^2) + O((n\lambda)^{-1})$ under model (2).

In the simulation study we are going to present in Section 3, we used a positive constant in place of the function $v(\cdot)$ in $\widehat{F}_{CD}(y)$, and local linear regression weights in the estimators involving nonparametric regressions. Thus, in the estimators $\widehat{F}_{DH}(y)$, $\widehat{F}_K(y)$ and $\widehat{F}_{CDW}(y)$ we replaced $w_j(x)$ by the jth component of the row-vector

$$\mathbf{w} := \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}_x, \tag{5}$$

where $\mathbf{e}_1^T := (1,0)$, $\mathbf{X}_x := [1, (x_i - x)]_{i\in s}$, and $\mathbf{W}_x := diag[K((x_i - x)/\lambda)]_{i\in s}$ with $\lambda > 0$ and with the Epanechnikov kernel function

$$K(u) := \frac{3}{4}(1-u^2) I(|u| \le 1), \quad u \in \mathbb{R}$$

in place of $K(\cdot)$.

2.2 Model-assisted estimators

Including design weights in the predictors for the unobserved indicator functions $I(y_i \leq y)$ yields fitted values g_i for use in *model-assisted* estimators of $F_N(y)$. Such fitted values may be incorporated either in a *generalized difference* estimator or in *model-calibrated* estimators. Rao, Kovar and Mantel [12] follow the former approach. They use

$$g_i := \left(\sum_{j\in s} d_j\right)^{-1} \sum_{j\in s} d_j I\left(\frac{y_j - x_j\widetilde{\beta}}{v(x_j)} \leq \frac{y - x_i\widetilde{\beta}}{v(x_i)}\right)$$

as fitted values for the unobserved $I(y_i \leq y)$'s (i.e. for $i \notin s$), and

$$g_i := \left(\sum_{j\in s} d_{ij}\right)^{-1} \sum_{j\in s} d_{ij} I\left(\frac{y_j - x_j\widetilde{\beta}}{v(x_j)} \leq \frac{y - x_i\widetilde{\beta}}{v(x_i)}\right)$$

as fitted values for the observed $I(y_i \leq y)$'s (i.e. for $i \in s$). In the definition of the fitted values g_i,

$$\widetilde{\beta} := \frac{\sum_{i\in s} d_i v^{-2}(x_i) x_i y_i}{\sum_{i\in s} d_i v^{-2}(x_i) x_i^2}$$

is an asymptotically design-unbiased estimator for

$$\beta_N := \frac{\sum_{i=1}^N v^{-2}(x_i) x_i y_i}{\sum_{i=1}^N v^{-2}(x_i) x_i^2},$$

the weighted least squares estimator for β in model (1) based on the whole population, and $d_{ij} := \frac{\pi_i}{\pi_{ij}}$ denotes the inverse of the conditional probability that unit j belongs to the sample given that unit i is present in the sample. Using these fitted values g_i in a *generalized difference* estimator yields

$$\widetilde{F}_{RKM}(y) := \frac{1}{N}\left\{\sum_{i\in s} d_i I(y_i \leq y) + \left(\sum_{i=1}^N g_i - \sum_{i\in s} d_i g_i\right)\right\}.$$

In the simulation study we are going to present in the following Section, we analyzed a slightly modified version of $\widetilde{F}_{RKM}(y)$, where the inverse conditional inclusion probabilities d_{ij} are replaced by the inverse (marginal) inclusion probabilities d_j. The resulting estimator will be denoted by $\widetilde{F}_{CD}(y)$ below, since it may be viewed as a *model-assisted* version of $\widehat{F}_{CD}(y)$. Note that in the case of Poisson sampling the definitions $\widetilde{F}_{CD}(y)$ and $\widetilde{F}_{RKM}(y)$ coincide, while in case of simple random without replacement sampling $\widetilde{F}_{RKM}(y)$ and $\widetilde{F}_{CD}(y)$ give rise estimates that are only slightly different if the population size N is large. As in $\widehat{F}_{CD}(y)$, we used a positive constant in place of the function $v(\cdot)$ in the estimator $\widetilde{F}_{CD}(y)$ in the simulations.

Further *generalized difference* estimators based on fitted values derived from the predictors $\widehat{G}_{DH,i}$ and $\widehat{p}_{K,i}$ are defined and analyzed in [7] for the case of simple random without replacement sampling. But while in [7] the fitted values g_i are based on leave-one-out estimators if $i \in s$, those used in the *model-assisted* estimators in the simulation study presented below are based on all sample observations regardless of whether $i \in s$ or $i \notin s$. The fitted values derived from $\widehat{G}_{DH,i}$ were computed by

$$g_i := \frac{1}{N}\sum_{j\in s} d_i I(y_j - \widetilde{\mu}(x_j) \leq y - \widetilde{\mu}(x_i)), \quad i = 1,2,\ldots,N. \tag{6}$$

In their definition

$$\widetilde{\mu}(x) := \sum_{i\in s} w_{\pi,i}(x) y_i \tag{7}$$

is a design-based version of $\widehat{\mu}(x)$, with $w_{\pi,i}(x)$ given by the components of the row-vector

$$\mathbf{w}_\pi := \mathbf{e}_1^T \left(\mathbf{X}_x^T \mathbf{W}_{\pi,x} \mathbf{X}_x\right)^{-1} \mathbf{X}_x^T \mathbf{W}_{\pi,x}, \tag{8}$$

where $\mathbf{W}_{\pi,x} := diag[d_i K((x_i - x)/\lambda)]_{i\in s}$. The *generalized difference* estimator based on the fitted values in (6) will be denoted by $\widetilde{F}_{DH}(y)$ below. Replacing the $w_i(\cdot)$'s by the design-weighted local linear regression weights $w_{\pi,i}(\cdot)$ in the definition of the predictors $\widehat{p}_{K,i}$, yields another set of fitted values g_i that we included in a *generalized difference* estimator denoted by $\widetilde{F}_K(y)$ below. Going one step further than in [7], we derived fitted values also from the predictor $\widehat{p}_{CDW,i}$ to define a *generalized difference* estimator that provides a *model-assisted* version of the estimator $\widehat{F}_{CDW}(y)$. The resulting estimator will be denoted by $\widetilde{F}_{CDW}(y)$. The fitted values g_i used in $\widetilde{F}_{CDW}(y)$ are given

$$g_i := g_{CD,i} + \sum_{j\in s} w_{\pi,j}(x_i)\left(g_{K,i} - g_{CD,i}\right), \quad i = 1,2,\ldots,N, \tag{9}$$

where $g_{CD,i}$ and $g_{K,i}$ denote the fitted values used in $\widetilde{F}_{CD}(y)$ and $\widetilde{F}_K(y)$.

Model-calibrated estimators for $F_N(y)$ are obtained by replacing the inverse inclusion probabilities d_i in the Horvitz-Thompson estimator $\widehat{F}_{HT}(y)$ by a set of sample weights ω_i. The sample weights ω_i are defined as the solution to a constrained optimization problem, where some function that measures the distance between the sample weights ω_i and the inverse inclusion probabilities d_i is minimized, under the constraint that

$$\sum_{i\in s} \omega_i g_i = \sum_{i=1}^{N} g_i \tag{10}$$

for some set of fitted values g_i. As for the distance function to be minimized, two popular choices are given by the Kullback-Leibler distance between the ω_i's and the d_i's (see [18, 17, 5, 10, 11]) and by some weighted chi-squared distance (see [6, 16, 13, 14]). The former choice leads to the class of pseudo empirical maximum likelihood estimators. One drawback of this class of estimators is that there exists no closed form solution for the corresponding weights ω_i. We therefore did not include

pseudo empirical maximum likelihood estimators for $F_N(y)$ in the simulation study. Their performance relative to other estimators included in the simulation study has been tested in another simulation study in [14]. On the other hand, minimizing the weighted chi-squared distance

$$\Phi_s := \sum_{i \in s} \frac{(\omega_i - d_i)^2}{d_i q_i}, \tag{11}$$

under the constraint in (10) does lead to a closed form solution for the sample weights ω_i. The solution is given by

$$\omega_i = d_i + \left(\sum_{j=1}^{N} g_j - \sum_{j \in s} d_j g_j \right) \frac{d_i q_i g_i}{\sum_{j \in s} d_j q_j g_j^2}, \quad i \in s.$$

Applying these weights to the observed indicator functions $I(y_i \leq y)$ yields the model calibrated estimator

$$\widetilde{F}^*_{MC}(y) := \frac{1}{N} \sum_{i \in s} \omega_i I(y_i \leq y)$$

for $F_N(y)$. It is worth noting that $\widetilde{F}^*_{MC}(y)$ may give rise to estimates that are not genuine distribution functions, since some sample weights ω_i might be negative, and since the weights ω_i depend on the fitted values g_i, which in turn depend on the y-value at which the distribution function $F_N(y)$ has to be estimated. [13, 14] propose a solution to these problems. In these papers it is suggested to consider a single set of sample weights ω_i to estimate $F_N(y)$ at all y-values. The set of sample weights ω_i should be obtained by considering instead of a single set of fitted values g_i, r sets of fitted values that do not depend on the y-value at which $F_N(y)$ has to be estimated. In particular, it is suggested to use the r sets of fitted values given by

$$g_{ij} := I(\widetilde{\mu}(x_i) \leq y^*_j), \quad i = 1,2,\ldots,N; \quad j = 1,2,\ldots,r,$$

where $y^*_1 < y^*_2 < \cdots < y^*_r$ are fixed, and $\widetilde{\mu}(\cdot)$ may be given either by

$$\widetilde{\mu}(x) := x\widetilde{\beta}, \quad x \in \mathbb{R},$$

under model (1), or by some design-weighted nonparametric estimator for the regression function $\mu(\cdot)$. In the simulation study we followed [14] and used (7) to compute the $\widetilde{\mu}(x_i)$'s in the fitted values g_{ij}. Still following [14], we set $r = 4$ and used the three quartiles of the $\widetilde{\mu}(x_i)$'s, $i = 1,2,\ldots,N$, in place of y^*_1, y^*_2, y^*_3, and $\max_{1 \leq i \leq N} \widetilde{\mu}(x_i)$ in place of y^*_4. The latter choice makes sure that the sample weights ω_i do sum to 1. Minimizing the chi-squared distance function (11) with constant weights q_i subject to the set of constraints

$$\sum_{i \in s} \omega_i \mathbf{g}_i = \mathbf{g}_N,$$

where $\mathbf{g}_i = (g_{i1}, g_{i2}, \ldots, g_{ir})^T$ and $\mathbf{g}_N = \sum_{i=1}^{N} \mathbf{g}_i$, yields a set of nonnegative sample weights ω_i as shown in [13, 14]. These sample weights are given by

$$\omega_i = d_i(1 + \gamma\, \mathbf{g}_i), \quad i \in s,$$

where

$$\gamma = \left(\mathbf{g}_N - \sum_{i \in s} d_i \mathbf{g}_i \right)^T \left(\sum_{i \in s} d_i \mathbf{g}_i \mathbf{g}_i^T \right)^{-1}$$

Since the sample weights ω_i sum to 1, the estimator

$$\widetilde{F}_{MC}(y) = \frac{1}{N} \sum_{i \in s} \omega_i I(y_i \leq y)$$

provides estimates that are genuine distribution functions with probability 1. Given its construction, the estimator $\widetilde{F}_{MC}(y)$ should be particularly efficient at y-values close to $y_1^*, y_2^*, \ldots, y_r^*$.

3 The double regression estimator

The estimator we propose in this work employs still another approach to compute predictions for the unobserved indicator functions $I(y_i \leq y)$. To accommodate super-population models like the one in (2) with possibly not identically distributed error components ε_i, we propose to employ two nonparametric regressions to estimate first the model-mean regression function $\mu(\cdot)$, and then, using the fitted residuals, the model-cdfs $G_i(\varepsilon) = P(\varepsilon_i \leq \varepsilon)$ of the error components. The resulting predictors for the unobserved indicator functions $I(y_i \leq y)$ are thus given by

$$\widehat{p}_{DR,i} := \sum_{j \in s} w_j(x_i) I\left(y_j - \widehat{\mu}(x_j) \leq y - \widehat{\mu}(x_i)\right), \quad i \notin s,$$

and the corresponding *model-based* estimator for $F_N(y)$ is

$$\widehat{F}_{DR}(y) := \frac{1}{N} \left\{ \sum_{i \in s} I(y_i \leq y) + \sum_{i \notin s} \widehat{p}_{DR,i} \right\}.$$

Because $\widehat{F}_{DR}(y)$ is based on two regressions we called it "Double Regression" estimator. It is worth noting that $\widehat{F}_{DR}(y)$ provides estimates that are genuine distribution functions with probability 1 if the regression weights $w_j(\cdot)$ are nonnegative and if they do sum to 1. In some situations it should be a good idea to employ different regression weights $w_j(\cdot)$ for estimating the regression function $\mu(\cdot)$ and the cdfs $G_i(\cdot)$

of the error components.[1] In particular, if the sample or a priori information support the hypothesis of a linear regression function $\mu(\cdot)$, there should be some gain in efficiency by using even linear regression instead of nonparametric regression to estimate $\mu(\cdot)$. In this way $\widehat{F}_{DR}(y)$ would yield perfect estimates in the scholastic limit case where $y_i := x_i\beta$ for all $i = 1,2,\ldots,N$. It is worth noting that this goal can also be achieved by using local linear regression to estimate the regression function $\mu(\cdot)$. On the other hand, if the sample or a priori information support the hypothesis of identically distributed error components, then one should simply put $w_j(x_i) = 1/n$ for all $j \in s$ and for all $i \notin s$ in the estimation of the $G_i(\cdot)$'s. In this case $\widehat{F}_{DR}(y)$ reduces obviously to $\widehat{F}_{DH}(y)$. Based on these considerations we should thus expect that: *(i)* $\widehat{F}_{DR}(y)$ is not as efficient as $\widehat{F}_{CD}(y)$ if the population follows model (1); *(ii)* $\widehat{F}_{DR}(y)$ is not as efficient as $\widehat{F}_{DH}(y)$ if model (2) with i.i.d. error components provides a good fit to the population. On the other hand, $\widehat{F}_{DR}(y)$ should be more efficient than both $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$ in populations generated from model (2) with not identically distributed error components because of model misspecification bias in $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$.

To hint at conditions under which $\widehat{F}_{DR}(y)$ should be more efficient than $\widehat{F}_{K}(y)$ and/or $\widehat{F}_{CDW}(y)$ we observe that the superpopulation model underlying the latter two estimators is slightly more general than the one that leads to the definition of $\widehat{F}_{DR}(y)$. In fact, model (3) allows the regression function $\mu(\cdot)$ even to not exist, and since $\widehat{F}_{DR}(y)$ is based on some estimate for $\mu(\cdot)$, we should expect quite erratic estimates from $\widehat{F}_{DR}(y)$ in populations generated from such models. But otherwise, especially when the population exhibits a strong regression relationship between the study variable Y and the auxiliary variable X, we should expect that $\widehat{F}_{DR}(y)$ is more efficient than $\widehat{F}_{K}(y)$, since $\widehat{F}_{DR}(y)$ exploits a separate estimate of $\mu(\cdot)$. For the same reason $\widehat{F}_{DR}(y)$ should be expected to be more efficient than $\widehat{F}_{CDW}(y)$ too, if there is a strong regression relationship and model (1) does not provide a good fit.

To accommodate sample designs with markedly nonconstant inclusion probabilities π_i, we propose also a *model-assisted* version of $\widehat{F}_{DR}(y)$ in form of a *generalized difference* estimator. The *model-assisted* version will be denoted by $\widetilde{F}_{DR}(y)$. The fitted values g_i to be used in $\widetilde{F}_{DR}(y)$ are obtained by incorporating the sample inclusion probabilities π_i in the regression weights $w_j(\cdot)$ in the predictors $\widehat{p}_{DR,i}$. Considerations about the efficiency of $\widetilde{F}_{DR}(y)$ in comparison to that of the *generalized difference* estimators $\widetilde{F}_{CD}(y)$, $\widetilde{F}_{DH}(y)$, $\widetilde{F}_{K}(y)$ and $\widetilde{F}_{CDW}(y)$ are similar to those already made above concerning the *model-based* versions of these estimators. As for comparison with $\widetilde{F}_{MC}(y)$, we should expect that $\widetilde{F}_{DR}(y)$ is about as efficient as $\widetilde{F}_{MC}(y)$ in case of a strong regression relationship, since both estimators exploit an estimate of $\mu(\cdot)$, and that $\widetilde{F}_{DR}(y)$ is more efficient than $\widetilde{F}_{MC}(y)$ in case of a weak regression relationship, since the fitted values used in $\widetilde{F}_{MC}(y)$ will be unreliable in the latter case, while those used in $\widetilde{F}_{DR}(y)$ will track the indicator functions $I(y_i \leq y)$

[1] In this case it is enough that the regression weights $w_j(\cdot)$ used for estimating the cdfs of the error components are nonnegative and that they sum to 1 to ensure that $\widehat{F}_{DR}(y)$ provides estimates that are genuine distribution functions with probability 1.

closer, because they are sensitive to both the fitted regression function and the fitted regression residuals.

In the simulation study presented in the next section we tested $\widehat{F}_{DR}(y)$ and $\widetilde{F}_{DR}(y)$ with the local linear regression weights defined in (5) and in (8), respectively. We also tried to use different bandwidths λ to estimate $\mu(\cdot)$ and the cdfs of the error components $G_i(\cdot)$ as will be explained in the next section.

4 Simulation Results

In this Section we report the results of a simulation study where the double regression estimators $\widehat{F}_{DR}(y)$ and $\widetilde{F}_{DR}(y)$ have been compared with the estimators recalled in Section 2. In the simulations we considered finite populations of size $N = 1000$ generated from the following superpopulation models:

$$\begin{aligned} M1 \qquad & y_i := 0.5x_i + \sigma\varepsilon_i \\ M2 \qquad & y_i := \log(x_i) + \sigma\varepsilon_i \\ M3 \qquad & y_i := \sqrt{x_i} + \sigma\varepsilon_i \\ M4 \qquad & y_i := -10(x_i - 0.5)^2 + \sigma\varepsilon_i \ . \end{aligned}$$

As for the error components ε_i, they were generated independently from either the Student t distribution with $\nu = 5$ df (identically distributed error components), or from shifted noncentral Student t distributions with $\nu = 5$ dgf and with noncentrality parameter given by $\zeta = 15x_i$ (not identically distributed error components). The shifts applied to the error component distributions in the latter case are aimed to make sure that their expectations are equal to zero. Note that, in the case of not identically distributed error components, the value of the auxiliary variable influences not only the scale of the error distributions but also their entire shape. As for σ, we considered two values: $\sigma = 0.1$ (strong regression relationship), and $\sigma = 0.3$ (weak regression relationship). The x-values taken on by the auxiliary variable were independently generated from the uniform distribution on $(0,1)$. The populations we considered are thus $4(\text{superpopulation models}) \times 2(\text{types of error components}) \times 2(\text{values of } \sigma) = 16$. The scatterplots of the 16 considered populations are shown in Figure 1 and in Figure 2.

To compare the performance of the estimators, we selected $B = 1000$ samples from each of the 16 populations by simple random without replacement sampling (srwors) with sample sizes $n = 50$ and $n = 100$, and by Poisson sampling with inclusion probabilities π_i proportional to $\sqrt{x_i}$ and expected sample sizes $n = 50$ and $n = 100$. As anticipated in Section 2 and Section 3, in all estimators involving nonparametric regression, local linear regression weights with Epanechnikov kernel function were used. The bandwidth λ was set either equal to $\lambda_s := 0.20$ (small bandwidth) or to $\lambda_l := 0.40$ (large bandwidth) for $n = 50$, and either equal to $\lambda_s := 0.15$ (small bandwidth) or to $\lambda_l := 0.30$ (large bandwidth) for $n = 100$. In $\widehat{F}_{DR}(y)$ and $\widetilde{F}_{DR}(y)$ we also tried to use two different bandwidths λ_μ and λ_G for estimating the

regression function $\mu(\cdot)$ and the cdfs $G_i(\cdot)$ of the error components, respectively. The combinations we tested are given by $(\lambda_\mu, \lambda_G) := (\lambda_s, \lambda_s)$; $(\lambda_\mu, \lambda_G) := (\lambda_s, \lambda_l)$; $(\lambda_\mu, \lambda_G) := (\lambda_l, \lambda_s)$; $(\lambda_\mu, \lambda_G) := (\lambda_l, \lambda_l)$.

For each considered estimator, we evaluated the estimation error with respect to $F_N(t_k)$ at $t_k = F_N^{-1}(k/20)$ for $k = 1, 2, \ldots, 19$, and we computed the simulated Relative Bias (RB) and the simulated Relative Mean Squared Error (RMSE). The simulated RB for estimator "$\bullet$" at the point t_k is defined by

$$RB_\bullet(k) = \frac{B_\bullet(k)}{(k/20)} \qquad k = 1, 2, ..., 19,$$

where $B_\bullet(k)$ denotes the simulated bias of estimator "$\bullet$" at the point t_k. The simulated RMSE for estimator "$\bullet$" at the point t_k is computed by taking $\widehat{F}_{HT}(y)$ as benchmark. It is defined by

$$RMSE_\bullet(k) = \frac{MSE_\bullet(k)}{MSE_{HT}(k)} \qquad k = 1, 2, ..., 19$$

where $MSE_\bullet(k)$ denotes the simulated MSE of estimator "$\bullet$" at the point t_k. Based on $RB_\bullet(k)$ and on $RMSE_\bullet(k)$, $k = 1, 2, \ldots, 19$, we finally computed the AVeraged Relative Bias (AVRB) and the AVeraged Relative Error (AVRE)

$$AVRB_\bullet = \frac{1}{19} \sum_{k=1}^{19} |RB_\bullet(k)|\,, \qquad AVRE = \frac{1}{19} \sum_{k=1}^{19} RMSE_\bullet(k)$$

to evaluate the overall performance of each estimator. Tables 1 to 8 report the values of both these indexes for each considered estimator, population and sample design. In these tables the notation identifying the estimators has been slightly modified in order to single out, if necessary, the results obtained with *small* and *large* bandwidth. In detail, we added an "*s*" or an "*l*" to the subscript of $\widehat{F}_\bullet(y)$ and $\widetilde{F}_\bullet(y)$ according to the bandwidth. In $\widehat{F}_{DR}(y)$ and $\widetilde{F}_{DR}(y)$ we added two subscripts "*ss*", "*sl*", "*ls*", or "*ll*", where the first one refers to the bandwidth used for estimating $\mu(\cdot)$, and the second one to the bandwidth for estimating the $G_i(\cdot)$'s. The AVRB and AVRE-values reported in Tables 1 to 4 and those referring to the estimators that do not involve local-linear regression with bandwidth λ_s in Tables 5 to 8 are based on $B = 1000$ samples, while those referring to the estimators that do involve local-linear regression with bandwidth λ_s in Tables 5 and 6 and in Tables 7 and 8 are based on 878 and 969 samples, respectively. This is caused by the fact that while under the srwors sample designs no sample occurred with not well-defined local-linear regression weights (which happens when there are too large holes in the sampled x-values), under the Poisson sample designs with expected sample sizes $n = 50$ and $n = 100$ there occurred respectively 122 and 31 samples where at some population x-value the local-linear regression weights $w_j(x)$ with bandwidth λ_s were not well-defined.

The simulation results reported in Tables 1 to 8 confirm that the estimators based on auxiliary information do often provide a large gain in efficiency with respect to

the Horvitz-Thompson estimator. In fact, the AVRE of the former estimators is usually smaller than 1 and it may reach values close to 0.1 in some cases. The gain in efficiency (as measured by the AVRE) seems less evident in the populations generated from the models with not identically distributed error components. Perhaps surprisingly, the AVRE of the *model-based* estimators tends to be smaller than that of the *model-assisted* ones not just under the srwors schemes, but also under the Poisson sampling schemes. In fact, under srwors there is, as expected, a large positive gap between the lowest AVRE among the *model-based* estimators, and the lowest AVRE among the *model-assisted* ones. Under the Poisson sampling schemes this gap reduces, but except for the population generated from model M1 with $\sigma = 0.3$, where $\widetilde{F}_{MC}(y)$ has lowest AVRE among all considered estimators when the expected sample size is $n = 50$, the smallest AVRE among the *model-based* estimators is always lower than the smallest AVRE among the *model-assisted* ones. This outcome is explained by the fact that the AVRB of the *model-based* estimators remains quite low under the Poisson sampling schemes as well, even though the inclusion probabilities π_i are proportional to $\sqrt{x_i}$. More variability in the inclusion probabilities would likely change this result.

Comparing the *model-based* estimators, we note that $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$ usually have lowest AVRE in the populations generated from the models with identically distributed error components, and that the new estimator $\widehat{F}_{DR}(y)$ has usually lowest AVRE in the other populations. But while the AVRE of $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$ is huge in the populations generated from models with not identically distributed error components, that of $\widehat{F}_{DR}(y)$ remains quite close to that $\widehat{F}_{CD}(y)$ and $\widehat{F}_{DH}(y)$ when the latter estimators are more precise. $\widehat{F}_{K}(y)$ and $\widehat{F}_{CDW}(y)$ exhibit also low AVRE-values in many instances. However, the AVRE of the latter estimators is usually larger than that of $\widehat{F}_{DR}(y)$. In favor of $\widehat{F}_{DR}(y)$ it is further worth noting that its AVRE is not as sensitive to the bandwidth as that of $\widehat{F}_{DH}(y)$, $\widehat{F}_{K}(y)$ and $\widehat{F}_{CDW}(y)$. The more pronounced sensitivity to the bandwidth of in particular $\widehat{F}_{K}(y)$ and $\widehat{F}_{CDW}(y)$ is exacerbated by the Poisson sampling schemes. This is probably due to the fact that $\widehat{F}_{K}(y)$ and $\widehat{F}_{CDW}(y)$ (in its bias correction term) do not take advantage of a separate estimate of the regression function, which makes the predictors used those estimators very variable if the bandwidth is small.

Similar considerations as those about the *model-based* estimators, hold for comparisons among the *generalized difference* estimators derived from $\widehat{F}_{CD}(y)$, $\widehat{F}_{DH}(y)$, $\widehat{F}_{K}(y)$, $\widehat{F}_{CDW}(y)$ and $\widehat{F}_{DR}(y)$. As for $\widetilde{F}_{MC}(y)$, the only *model-calibrated* estimator included in the simulation study, it usually exhibits a larger AVRE than the *generalized difference* estimators under the srwors schemes. This may be the price to be paid by the estimator for providing always estimates that are genuine distribution functions. However, under the Poisson sampling schemes $\widetilde{F}_{MC}(y)$ has in many cases lowest AVRE among the *model-assisted* estimators when the population is generated from models with not identically distributed error components. In one instance, as already mentioned above, $\widetilde{F}_{MC}(y)$ is even the estimator with lowest AVRE among all estimators included in the simulation study. The competitiveness of $\widetilde{F}_{MC}(y)$ under the Poisson sampling schemes is due to fact that $\widetilde{F}_{MC}(y)$ is based on four sets of fitted values, while the other estimators are based on a single set of predictions or

fitted values. In fact, under the Poisson sampling schemes the predictions and fitted values for $I(y_i \leq y)$ at low x-values are very variable, while the high variability of the set of fitted values corresponding to y_1^* in $\widetilde{F}_{MC}(y)$ is dampened by lower variability of the other three sets of fitted values corresponding to y_2^*, y_3^* and y_4^*. The AVRE of $\widetilde{F}_{DR}(y)$ is however very close to that $\widetilde{F}_{MC}(y)$ when the latter estimator performs better, unless the wrong bandwidth combination is used in $\widetilde{F}_{DR}(y)$.

5 Conclusions

In this work we proposed a new estimator for a finite population distribution function based on a very general superpopulation model that allows for nonlinear regression functions and possibly not identically distributed error components. The new estimator employs two local linear regressions to estimate first the regression function and then the cdfs of the error components. We therefore call it "Double Regression" estimator (DR-estimator). We proposed a *model-based* as well as a *model-assisted* version of the new estimator, the latter in the form of *generalized difference* estimator. We hint at conditions under which the new estimator should outperform some well-known estimators from the literature and tested the efficiency of the new estimator in a simulation study involving several populations and two sample designs: simple random without replacement sampling and Poisson sampling with nonconstant inclusion probabilites. The simulation results show that the DR-estimator looses little efficiency with respect to the Chambers and Dunstan estimator [4] and with respect to the nonlinear version Chambers and Dunstan estimator proposed in [7] in populations generated from the more restrictive models underlying the definitions of the latter estimators. On the other hand, the DR-estimator is clearly more efficient in populations with nonlinear regression functions and sparse regression residuals. The DR-estimator also outperforms the Kuo [9] and by Chambers, Dunstan and Wehrly [3] estimators in all considered populations and under both sample designs. Among the considered settings, the model-calibrated estimator proposed in [14] turned out to be a little more efficient than the DR-estimator just in one case, while being far less efficient in several other cases.

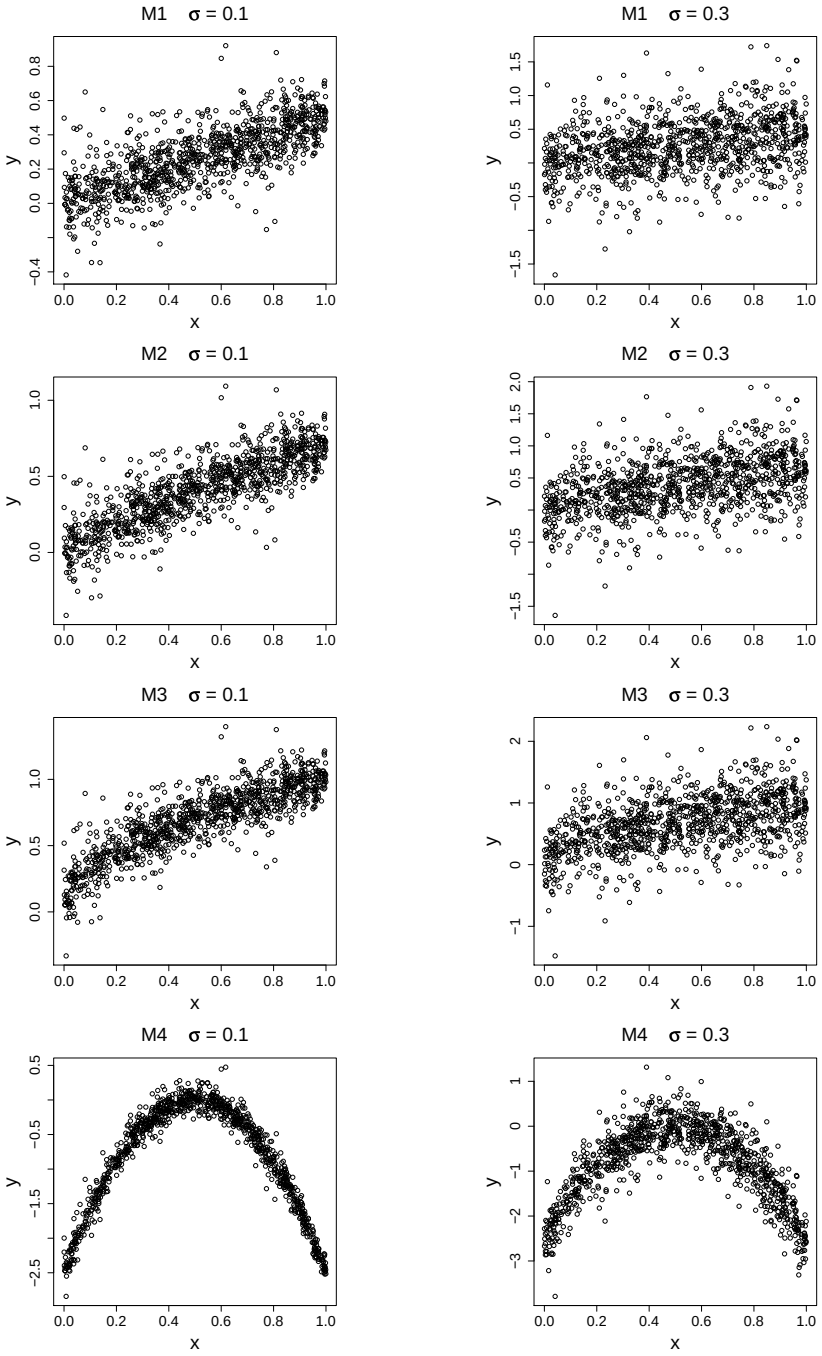

Fig. 1 Finite populations generated from models M1-M4 with $\sigma = 0.1$ and $\sigma = 0.3$ and identically distributed error components.

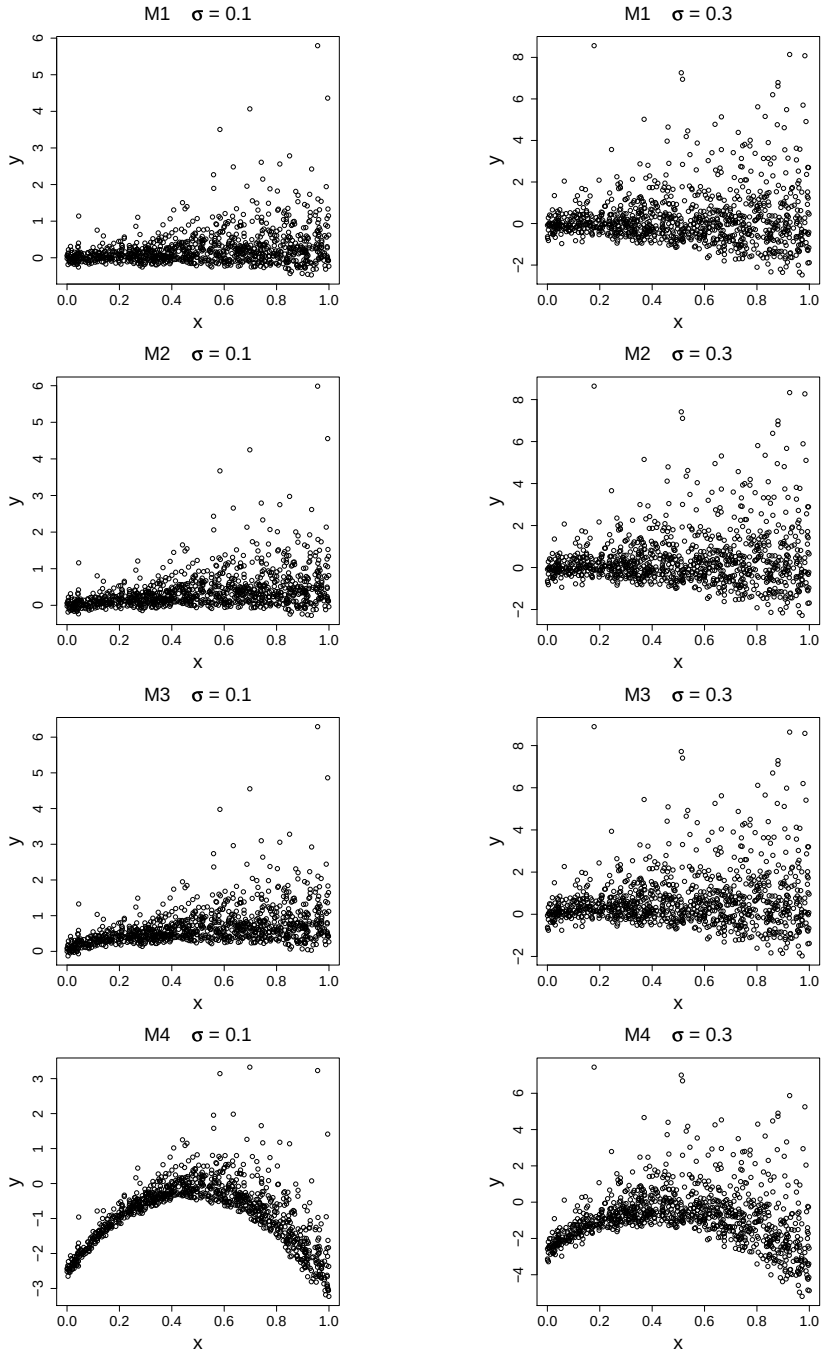

Fig. 2 Finite populations generated from models M1-M4 with $\sigma = 0.1$ and $\sigma = 0.3$ and not-identically distributed error components.

Table 1 Populations generated from the models with identically distributed error components. Simulated AVRB and AVRE in case of simple random without replacement sampling. Sample size $n = 50$.

Model	$\sigma = 0.1$								$\sigma = 0.3$							
	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0072	1.0000	0.0032	1.0000	0.0060	1.0000	0.0039	1.0000	0.0034	1.0000	0.0041	1.0000	0.0049	1.0000	0.0067	1.0000
$\widehat{F}_{CD}$	0.0180	0.3719	0.0294	0.3456	0.0752	0.4088	0.0268	1.0004	0.0284	0.7429	0.0172	0.6569	0.0293	0.6090	0.0223	0.9238
$\widehat{F}_{DHs}$	0.0208	0.4627	0.0161	0.3657	0.0134	0.3218	0.0359	0.1562	0.0280	0.7995	0.0184	0.6893	0.0223	0.6783	0.0309	0.3466
$\widehat{F}_{DHl}$	0.0198	0.4151	0.0141	0.3324	0.0233	0.2831	0.1683	1.2266	0.0293	0.7697	0.0158	0.6683	0.0248	0.6351	0.1264	0.7535
$\widehat{F}_{Ks}$	0.0072	0.7716	0.0250	0.6640	0.0272	0.6022	0.0844	0.6254	0.0031	1.0857	0.0126	1.0279	0.0096	0.9940	0.0742	0.7594
$\widehat{F}_{Kl}$	0.0441	0.8342	0.0769	0.8209	0.0816	0.7703	0.1930	1.7554	0.0139	1.0751	0.0283	1.0120	0.0317	0.9957	0.1761	1.6428
$\widehat{F}_{CDWs}$	0.0086	0.7675	0.0084	0.6458	0.0137	0.5885	0.0840	0.6271	0.0043	1.0891	0.0078	1.0276	0.0061	0.9966	0.0737	0.7615
$\widehat{F}_{CDWl}$	0.0078	0.7401	0.0194	0.6298	0.0296	0.5825	0.1918	1.7644	0.0073	1.0772	0.0164	1.0025	0.0145	0.9792	0.1748	1.6498
$\widehat{F}_{DRs}$	0.0103	0.6231	0.0124	0.4741	0.0095	0.4245	0.0114	0.1415	0.0153	0.9521	0.0126	0.8674	0.0088	0.8583	0.0117	0.3765
$\widehat{F}_{DRl}$	0.0144	0.5280	0.0149	0.3953	0.0137	0.3398	0.0533	0.3085	0.0163	0.9115	0.0120	0.8116	0.0099	0.7930	0.0436	0.3965
$\widehat{F}_{DRsl}$	0.0140	0.5597	0.0128	0.4214	0.0121	0.3740	0.0166	0.1294	0.0201	0.9141	0.0145	0.8058	0.0143	0.8059	0.0217	0.3430
$\widehat{F}_{DRls}$	0.0105	0.6165	0.0123	0.4711	0.0076	0.4097	0.0201	0.1419	0.0131	0.9843	0.0095	0.9023	0.0075	0.8742	0.0152	0.3662
$\widetilde{F}_{CDs}$	0.0060	0.6660	0.0031	0.5664	0.0088	0.5260	0.0167	1.0536	0.0043	0.9423	0.0045	0.8912	0.0066	0.8618	0.0181	1.0364
$\widetilde{F}_{DHs}$	0.0061	0.7034	0.0025	0.5695	0.0062	0.5303	0.0083	0.2635	0.0033	0.9965	0.0051	0.9109	0.0037	0.9120	0.0073	0.5248
$\widetilde{F}_{DHl}$	0.0063	0.6787	0.0029	0.5584	0.0065	0.5147	0.0155	0.3652	0.0043	0.9606	0.0045	0.8927	0.0052	0.8770	0.0140	0.5409
$\widetilde{F}_{Ks}$	0.0064	0.7445	0.0044	0.6135	0.0048	0.5687	0.0099	0.4144	0.0031	1.0450	0.0044	0.9729	0.0055	0.9590	0.0057	0.5973
$\widetilde{F}_{Kl}$	0.0062	0.7066	0.0048	0.5932	0.0062	0.5523	0.0108	0.5304	0.0044	0.9823	0.0040	0.9238	0.0056	0.9024	0.0105	0.6317
$\widetilde{F}_{CDWs}$	0.0226	0.7383	0.0406	0.6623	0.0401	0.6016	0.0918	1.2304	0.0170	1.0441	0.0311	1.0114	0.0258	0.9647	0.0798	1.1977
$\widetilde{F}_{CDWl}$	0.0479	0.7990	0.0820	0.8072	0.0865	0.7612	0.2087	2.4276	0.0186	1.0775	0.0474	1.0651	0.0438	1.0190	0.1915	2.2405
$\widetilde{F}_{DRs}$	0.0070	0.7312	0.0032	0.5862	0.0066	0.5485	0.0040	0.2546	0.0030	1.0325	0.0052	0.9519	0.0040	0.9470	0.0053	0.5310
$\widetilde{F}_{DRl}$	0.0066	0.6879	0.0032	0.5648	0.0065	0.5182	0.0100	0.2762	0.0044	0.9806	0.0044	0.9202	0.0052	0.8945	0.0091	0.5112
$\widetilde{F}_{DRsl}$	0.0068	0.7026	0.0030	0.5712	0.0066	0.5280	0.0066	0.2551	0.0032	0.9996	0.0047	0.9222	0.0044	0.9145	0.0068	0.5176
$\widetilde{F}_{DRls}$	0.0069	0.7238	0.0034	0.5868	0.0063	0.5432	0.0040	0.2529	0.0037	1.0335	0.0047	0.9599	0.0055	0.9400	0.0042	0.5235
$\widetilde{F}_{MCs}$	0.0070	0.7525	0.0031	0.6620	0.0064	0.6195	0.0042	0.4271	0.0032	1.0468	0.0046	0.9726	0.0049	0.9642	0.0036	0.6064
$\widetilde{F}_{MCl}$	0.0070	0.7360	0.0027	0.6576	0.0060	0.6088	0.0059	0.4335	0.0040	1.0199	0.0040	0.9850	0.0055	0.9287	0.0059	0.6070

Table 2 Populations generated from the models with not identically distributed error components. Simulated AVRB and AVRE in case of simple random without replacement sampling. Sample size $n = 50$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0030	1.0000	0.0023	1.0000	0.0026	1.0000	0.0025	1.0000	0.0051	1.0000	0.0043	1.0000	0.0051	1.0000	0.0039	1.0000
$\widehat{F}_{CD}$	0.1613	1.6161	0.1717	1.6106	0.1556	1.4045	0.0220	0.9286	0.0684	1.2699	0.0558	1.2576	0.1000	1.3594	0.0114	0.9960
$\widehat{F}_{DHs}$	0.1722	1.7296	0.1825	1.7658	0.1765	1.6621	0.0321	0.3620	0.0962	1.3431	0.0676	1.2089	0.1297	1.4874	0.0359	0.7691
$\widehat{F}_{DHl}$	0.1647	1.6468	0.1781	1.6969	0.1715	1.5905	0.1353	0.8076	0.0799	1.2802	0.0588	1.2165	0.1149	1.4044	0.0308	0.7110
$\widehat{F}_{Ks}$	0.0120	1.1286	0.0205	1.0673	0.0349	0.9816	0.0807	0.7564	0.0116	1.1409	0.0047	1.1010	0.0160	1.1478	0.0592	0.9525
$\widehat{F}_{Kl}$	0.0143	1.0646	0.0414	1.0463	0.0696	1.0331	0.1814	1.6555	0.0184	1.0796	0.0151	1.0686	0.0184	1.0928	0.1346	1.3716
$\widehat{F}_{CDWs}$	0.0081	1.1321	0.0134	1.0667	0.0264	0.9734	0.0801	0.7579	0.0109	1.1435	0.0041	1.1049	0.0143	1.1507	0.0587	0.9543
$\widehat{F}_{CDWl}$	0.0104	1.0797	0.0205	1.0352	0.0426	0.9815	0.1800	1.6603	0.0179	1.0894	0.0151	1.0825	0.0149	1.1040	0.1332	1.3750
$\widehat{F}_{DRs}$	0.0299	0.9610	0.0250	0.8770	0.0311	0.7974	0.0169	0.3533	0.0326	1.0141	0.0118	0.9489	0.0317	1.0014	0.0166	0.6837
$\widehat{F}_{DRl}$	0.0410	0.9287	0.0388	0.8374	0.0455	0.7806	0.0527	0.3857	0.0403	1.0028	0.0235	0.9542	0.0392	0.9780	0.0251	0.6273
$\widehat{F}_{DRsl}$	0.0523	0.9567	0.0491	0.8714	0.0550	0.8135	0.0263	0.3256	0.0548	1.0273	0.0279	0.9359	0.0537	1.0018	0.0219	0.6420
$\widehat{F}_{DRls}$	0.0196	0.9995	0.0162	0.9039	0.0249	0.8247	0.0209	0.3582	0.0230	1.0639	0.0068	1.0187	0.0216	1.0493	0.0200	0.7011
$\widetilde{F}_{CDs}$	0.0063	0.9877	0.0048	0.9115	0.0060	0.8452	0.0146	1.0363	0.0066	1.0219	0.0050	1.0319	0.0074	1.0212	0.0132	1.0124
$\widetilde{F}_{DHs}$	0.0080	1.0544	0.0055	0.9735	0.0041	0.8887	0.0032	0.5130	0.0100	1.0908	0.0064	1.0546	0.0101	1.0906	0.0047	0.8391
$\widetilde{F}_{DHl}$	0.0064	1.0097	0.0037	0.9312	0.0036	0.8527	0.0097	0.5350	0.0075	1.0456	0.0044	1.0418	0.0078	1.0463	0.0029	0.8151
$\widetilde{F}_{Ks}$	0.0042	1.0668	0.0026	0.9996	0.0040	0.9093	0.0053	0.5754	0.0068	1.0866	0.0053	1.0439	0.0065	1.0870	0.0036	0.8418
$\widetilde{F}_{Kl}$	0.0043	0.9930	0.0032	0.9284	0.0044	0.8535	0.0065	0.6216	0.0065	1.0149	0.0049	0.9979	0.0065	1.0170	0.0017	0.8099
$\widetilde{F}_{CDWs}$	0.0290	1.1282	0.0372	1.0557	0.0516	0.9996	0.0882	1.2295	0.0271	1.1575	0.0249	1.1597	0.0315	1.1656	0.0682	1.1906
$\widetilde{F}_{CDWl}$	0.0328	1.1418	0.0608	1.1226	0.0906	1.1414	0.1973	2.2733	0.0281	1.1696	0.0260	1.1709	0.0359	1.1900	0.1507	1.7911
$\widetilde{F}_{DRs}$	0.0084	1.0535	0.0057	0.9812	0.0045	0.8915	0.0026	0.4978	0.0105	1.0760	0.0056	1.0305	0.0112	1.0755	0.0052	0.8129
$\widetilde{F}_{DRl}$	0.0053	0.9924	0.0031	0.9229	0.0040	0.8439	0.0067	0.4816	0.0079	1.0179	0.0057	1.0020	0.0079	1.0198	0.0025	0.7718
$\widetilde{F}_{DRsl}$	0.0075	1.0229	0.0047	0.9515	0.0038	0.8662	0.0043	0.4873	0.0097	1.0470	0.0058	1.0058	0.0097	1.0480	0.0038	0.7911
$\widetilde{F}_{DRls}$	0.0046	1.0482	0.0026	0.9770	0.0036	0.8887	0.0032	0.4962	0.0084	1.0742	0.0058	1.0391	0.0080	1.0746	0.0041	0.8077
$\widetilde{F}_{MCs}$	0.0041	1.0403	0.0030	1.0069	0.0031	0.9588	0.0024	0.6178	0.0068	1.1000	0.0050	1.0670	0.0065	1.0909	0.0042	0.8862
$\widetilde{F}_{MCl}$	0.0048	1.0262	0.0034	0.9737	0.0041	0.9173	0.0034	0.6084	0.0052	1.0638	0.0042	1.0559	0.0057	1.0582	0.0024	0.8737

Table 3 Populations generated from the models with identically distributed error components. Simulated AVRB and AVRE in case of simple random without replacement sampling. Sample size $n = 100$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0048	1.0000	0.0054	1.0000	0.0053	1.0000	0.0020	1.0000	0.0024	1.0000	0.0029	1.0000	0.0034	1.0000	0.0026	1.0000
$\widehat{F}_{CD}$	0.0156	0.4096	0.0281	0.3259	0.0722	0.5936	0.0147	1.0159	0.0254	0.7805	0.0275	0.6937	0.0264	0.6444	0.0124	0.9374
$\widehat{F}_{DHs}$	0.0158	0.4796	0.0140	0.3875	0.0103	0.3304	0.0143	0.1299	0.0242	0.8039	0.0247	0.7253	0.0186	0.6661	0.0184	0.3367
$\widehat{F}_{DHl}$	0.0160	0.4587	0.0132	0.3576	0.0153	0.3050	0.0917	0.7068	0.0252	0.8031	0.0261	0.7179	0.0196	0.6520	0.0685	0.5190
$\widehat{F}_{Ks}$	0.0047	0.7461	0.0082	0.6372	0.0187	0.5770	0.0598	0.5152	0.0034	1.0395	0.0045	0.9888	0.0075	0.9565	0.0503	0.6720
$\widehat{F}_{Kl}$	0.0265	0.7931	0.0397	0.7297	0.0547	0.7203	0.1367	1.4606	0.0093	1.0406	0.0144	0.9926	0.0226	0.9701	0.1282	1.4510
$\widehat{F}_{CDWs}$	0.0059	0.7461	0.0037	0.6351	0.0116	0.5715	0.0596	0.5162	0.0043	1.0417	0.0056	0.9918	0.0056	0.9586	0.0501	0.6734
$\widehat{F}_{CDWl}$	0.0040	0.7371	0.0086	0.6338	0.0257	0.5985	0.1362	1.4632	0.0051	1.0426	0.0070	0.9896	0.0148	0.9607	0.1278	1.4533
$\widehat{F}_{DRss}$	0.0075	0.6323	0.0094	0.5041	0.0075	0.4309	0.0105	0.1381	0.0126	0.9469	0.0112	0.8958	0.0072	0.8481	0.0097	0.3610
$\widehat{F}_{DRll}$	0.0127	0.5643	0.0105	0.4321	0.0086	0.3635	0.0163	0.1459	0.0118	0.9193	0.0098	0.8530	0.0093	0.7963	0.0186	0.3323
$\widehat{F}_{DRsl}$	0.0119	0.5801	0.0104	0.4528	0.0092	0.3857	0.0080	0.1277	0.0149	0.9106	0.0121	0.8493	0.0078	0.7954	0.0117	0.3333
$\widehat{F}_{DRls}$	0.0081	0.6305	0.0101	0.4988	0.0060	0.4223	0.0135	0.1340	0.0097	0.9693	0.0103	0.9117	0.0061	0.8593	0.0116	0.3532
$\widetilde{F}_{CD}$	0.0031	0.6938	0.0044	0.5884	0.0046	0.5501	0.0056	1.0400	0.0018	0.9589	0.0020	0.9148	0.0025	0.8867	0.0048	1.0279
$\widetilde{F}_{DHs}$	0.0032	0.7131	0.0040	0.6045	0.0033	0.5381	0.0032	0.2482	0.0023	0.9915	0.0020	0.9480	0.0024	0.9141	0.0030	0.5085
$\widetilde{F}_{DHl}$	0.0034	0.7031	0.0040	0.5956	0.0034	0.5315	0.0061	0.2871	0.0016	0.9767	0.0020	0.9331	0.0025	0.9015	0.0043	0.5100
$\widetilde{F}_{Ks}$	0.0038	0.7302	0.0033	0.6212	0.0033	0.5541	0.0040	0.3327	0.0018	1.0162	0.0025	0.9679	0.0031	0.9347	0.0029	0.5396
$\widetilde{F}_{Kl}$	0.0032	0.7154	0.0034	0.6100	0.0032	0.5504	0.0048	0.4413	0.0016	0.9891	0.0018	0.9435	0.0024	0.9117	0.0030	0.5714
$\widetilde{F}_{CDWs}$	0.0136	0.7392	0.0159	0.6281	0.0265	0.6000	0.0685	1.2162	0.0105	1.0266	0.0119	0.9815	0.0169	0.9576	0.0579	1.1750
$\widetilde{F}_{CDWl}$	0.0312	0.7880	0.0451	0.7292	0.0615	0.7608	0.1552	2.2741	0.0143	1.0474	0.0215	1.0092	0.0324	1.0032	0.1455	2.1380
$\widetilde{F}_{DRss}$	0.0035	0.7271	0.0034	0.6135	0.0032	0.5457	0.0026	0.2421	0.0018	1.0152	0.0023	0.9680	0.0030	0.9327	0.0034	0.5082
$\widetilde{F}_{DRll}$	0.0034	0.7079	0.0039	0.5984	0.0036	0.5326	0.0034	0.2483	0.0016	0.9878	0.0019	0.9435	0.0025	0.9096	0.0036	0.5002
$\widetilde{F}_{DRsl}$	0.0035	0.7139	0.0038	0.6040	0.0035	0.5370	0.0026	0.2435	0.0017	0.9939	0.0021	0.9493	0.0028	0.9152	0.0034	0.5028
$\widetilde{F}_{DRls}$	0.0037	0.7215	0.0035	0.6094	0.0033	0.5422	0.0026	0.2417	0.0020	1.0108	0.0027	0.9630	0.0032	0.9268	0.0031	0.5054
$\widetilde{F}_{MCs}$	0.0040	0.7377	0.0044	0.6410	0.0040	0.6043	0.0020	0.4009	0.0017	1.0048	0.0020	0.9543	0.0027	0.9309	0.0026	0.5805
$\widetilde{F}_{MCl}$	0.0038	0.7280	0.0044	0.6366	0.0040	0.5999	0.0035	0.4051	0.0017	0.9896	0.0023	0.9462	0.0028	0.9200	0.0034	0.5796

Table 4 Populations generated from the models with not identically distributed error components. Simulated AVRB and AVRE in case of simple random without replacement sampling. Sample size $n = 100$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0016	1.0000	0.0028	1.0000	0.0027	1.0000	0.0030	1.0000	0.0019	1.0000	0.0019	1.0000	0.0021	1.0000	0.0024	1.0000
$\widehat{F}_{CD}$	0.1539	2.3302	0.1647	2.4249	0.1489	2.0704	0.0120	0.9412	0.0587	1.3873	0.0781	1.5036	0.0917	1.6165	0.0081	1.0310
$\widehat{F}_{DHs}$	0.1621	2.4740	0.1749	2.6568	0.1690	2.5028	0.0203	0.3315	0.0817	1.4655	0.1011	1.6342	0.1161	1.8016	0.0337	0.7411
$\widehat{F}_{DHl}$	0.1554	2.3427	0.1702	2.5530	0.1659	2.4323	0.0773	0.5906	0.0694	1.3799	0.0899	1.5332	0.1061	1.6920	0.0263	0.6734
$\widehat{F}_{Ks}$	0.0098	1.0447	0.0137	0.9861	0.0247	0.9027	0.0556	0.6435	0.0098	1.0393	0.0102	1.0450	0.0127	1.0515	0.0397	0.8424
$\widehat{F}_{Kl}$	0.0158	1.0495	0.0336	1.0328	0.0552	1.0237	0.1310	1.4611	0.0124	1.0410	0.0144	1.0522	0.0197	1.0691	0.1000	1.3147
$\widehat{F}_{CDWs}$	0.0074	1.0483	0.0096	0.9879	0.0199	0.8991	0.0555	0.6449	0.0094	1.0416	0.0096	1.0479	0.0120	1.0549	0.0395	0.8432
$\widehat{F}_{CDWl}$	0.0085	1.0642	0.0195	1.0249	0.0385	0.9775	0.1306	1.4630	0.0116	1.0489	0.0126	1.0621	0.0172	1.0800	0.0995	1.3166
$\widehat{F}_{DRss}$	0.0174	0.9161	0.0136	0.8333	0.0181	0.7495	0.0152	0.3305	0.0202	0.9448	0.0187	0.9376	0.0193	0.9409	0.0141	0.6397
$\widehat{F}_{DRll}$	0.0273	0.9157	0.0261	0.8220	0.0344	0.7654	0.0293	0.3195	0.0245	0.9631	0.0245	0.9545	0.0250	0.9526	0.0213	0.6056
$\widehat{F}_{DRsl}$	0.0378	0.9387	0.0357	0.8484	0.0408	0.7839	0.0181	0.3082	0.0356	0.9654	0.0352	0.9538	0.0363	0.9583	0.0182	0.6123
$\widehat{F}_{DRls}$	0.0114	0.9391	0.0090	0.8502	0.0155	0.7635	0.0172	0.3291	0.0147	0.9780	0.0126	0.9743	0.0143	0.9738	0.0147	0.6429
$\widetilde{F}_{CD}$	0.0025	1.0202	0.0032	0.9439	0.0027	0.8748	0.0037	1.0219	0.0017	1.0366	0.0020	1.0420	0.0020	1.0436	0.0035	1.0142
$\widetilde{F}_{DHs}$	0.0028	1.0346	0.0027	0.9549	0.0026	0.8652	0.0042	0.4879	0.0033	1.0500	0.0036	1.0550	0.0037	1.0573	0.0025	0.7844
$\widetilde{F}_{DHl}$	0.0026	1.0235	0.0029	0.9457	0.0025	0.8594	0.0063	0.4962	0.0025	1.0385	0.0028	1.0437	0.0030	1.0464	0.0028	0.7774
$\widetilde{F}_{Ks}$	0.0022	1.0094	0.0023	0.9494	0.0022	0.8591	0.0047	0.5019	0.0035	1.0076	0.0035	1.0136	0.0035	1.0162	0.0034	0.7634
$\widetilde{F}_{Kl}$	0.0016	0.9901	0.0025	0.9309	0.0019	0.8483	0.0056	0.5495	0.0027	0.9900	0.0028	0.9966	0.0028	0.9997	0.0035	0.7636
$\widetilde{F}_{CDWs}$	0.0180	1.1127	0.0219	1.0345	0.0337	0.9768	0.0635	1.1935	0.0173	1.1269	0.0179	1.1325	0.0204	1.1402	0.0474	1.1548
$\widetilde{F}_{CDWl}$	0.0255	1.1522	0.0448	1.1299	0.0690	1.1623	0.1486	2.1784	0.0209	1.1592	0.0235	1.1730	0.0290	1.1954	0.1150	1.7868
$\widetilde{F}_{DRss}$	0.0028	1.0106	0.0024	0.9456	0.0027	0.8535	0.0043	0.4652	0.0039	1.0089	0.0042	1.0164	0.0042	1.0184	0.0027	0.7526
$\widetilde{F}_{DRll}$	0.0023	0.9868	0.0025	0.9267	0.0022	0.8403	0.0055	0.4620	0.0035	0.9900	0.0037	0.9959	0.0036	0.9990	0.0033	0.7406
$\widetilde{F}_{DRsl}$	0.0027	0.9986	0.0025	0.9352	0.0025	0.8457	0.0046	0.4645	0.0036	1.0007	0.0039	1.0077	0.0039	1.0095	0.0027	0.7480
$\widetilde{F}_{DRls}$	0.0024	1.0025	0.0021	0.9410	0.0023	0.8486	0.0045	0.4617	0.0039	1.0027	0.0041	1.0085	0.0040	1.0112	0.0025	0.7470
$\widetilde{F}_{MCs}$	0.0017	1.0069	0.0026	0.9577	0.0026	0.9038	0.0037	0.5740	0.0027	1.0503	0.0024	1.0388	0.0025	1.0318	0.0027	0.8215
$\widetilde{F}_{MCl}$	0.0016	0.9889	0.0027	0.9463	0.0025	0.8971	0.0043	0.5756	0.0028	1.0311	0.0025	1.0303	0.0025	1.0241	0.0031	0.8234

Table 5 Populations generated from the models with identically distributed error components. Simulated AVRB and AVRE in case of Poisson sampling with inclusion probabilities proportional to $\sqrt{x_i}$ and expected sample size $n = 50$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.2220	1.0000	0.2573	1.0000	0.2799	1.0000	0.0364	1.0000	0.0919	1.0000	0.1290	1.0000	0.1554	1.0000	0.0255	1.0000
$\widehat{F}_{CD}$	0.0203	0.1981	0.0475	0.1444	0.0923	0.1794	0.1667	3.0905	0.0221	0.6634	0.0299	0.5110	0.0432	0.4452	0.1495	2.2994
$\widehat{F}_{DHs}$	0.0403	0.2926	0.0289	0.2056	0.0303	0.1600	0.0399	0.1881	0.0208	0.7675	0.0240	0.6140	0.0295	0.5292	0.0274	0.4177
$\widehat{F}_{DHl}$	0.0303	0.2602	0.0216	0.1728	0.0309	0.1376	0.1699	1.1401	0.0190	0.7318	0.0216	0.5719	0.0248	0.4855	0.1175	0.6900
$\widehat{F}_{Ks}$	0.0186	2.2184	0.0275	1.7461	0.0199	1.0072	0.0553	1.0803	0.0134	4.2479	0.0168	3.6051	0.0126	3.0736	0.0589	2.1340
$\widehat{F}_{Kl}$	0.0362	0.5791	0.0500	0.4547	0.0580	0.3688	0.1257	1.7957	0.0087	1.2241	0.0130	1.0187	0.0119	0.8757	0.1211	1.6806
$\widehat{F}_{CDWs}$	0.0144	2.2019	0.0192	1.7364	0.0140	0.9988	0.0536	1.0706	0.0114	4.2486	0.0135	3.5900	0.0096	3.0593	0.0561	2.1311
$\widehat{F}_{CDWl}$	0.0061	0.5305	0.0104	0.3914	0.0269	0.3092	0.1208	1.8836	0.0065	1.2209	0.0048	1.0074	0.0071	0.8655	0.1140	1.7326
$\widehat{F}_{DRs}$	0.0236	0.4205	0.0205	0.2967	0.0252	0.2247	0.0122	0.2164	0.0173	1.0215	0.0152	0.8391	0.0177	0.7174	0.0172	0.5264
$\widehat{F}_{DRl}$	0.0192	0.3700	0.0144	0.2411	0.0259	0.1809	0.0578	0.3540	0.0132	0.9842	0.0099	0.7924	0.0131	0.6702	0.0419	0.4800
$\widehat{F}_{DRsl}$	0.0192	0.3538	0.0190	0.2429	0.0253	0.1859	0.0197	0.1598	0.0138	0.9051	0.0118	0.7296	0.0155	0.6223	0.0217	0.4286
$\widehat{F}_{DRls}$	0.0203	2.0509	0.0123	1.6025	0.0114	0.9682	0.0163	0.6291	0.0118	4.0313	0.0154	3.6233	0.0137	2.9631	0.0201	2.3122
$\widetilde{F}_{CDs}$	0.2077	0.8919	0.2414	0.8782	0.2604	0.8464	0.0496	1.0401	0.0854	0.9558	0.1193	0.9315	0.1429	0.9101	0.0367	1.0139
$\widetilde{F}_{DHs}$	0.0214	0.4011	0.0157	0.2973	0.0203	0.2418	0.0128	0.2920	0.0213	0.9437	0.0231	0.7750	0.0208	0.6695	0.0144	0.6250
$\widetilde{F}_{DHl}$	0.0125	0.3782	0.0082	0.2779	0.0216	0.2309	0.0280	0.4060	0.0125	0.8990	0.0138	0.7281	0.0140	0.6291	0.0199	0.6178
$\widetilde{F}_{Ks}$	0.0118	2.1745	0.0123	1.7075	0.0126	0.9817	0.0428	0.9318	0.0099	4.1725	0.0110	3.5408	0.0062	3.0209	0.0250	1.9698
$\widetilde{F}_{Kl}$	0.0079	0.4940	0.0105	0.3689	0.0249	0.2985	0.0484	0.8451	0.0062	1.1189	0.0050	0.9262	0.0117	0.7970	0.0358	0.9117
$\widetilde{F}_{CDWs}$	0.0117	2.1735	0.0121	1.7073	0.0124	0.9810	0.0428	0.9317	0.0098	4.1737	0.0109	3.5342	0.0063	3.0208	0.0250	1.9701
$\widetilde{F}_{CDWl}$	0.0079	0.4932	0.0105	0.3680	0.0248	0.2978	0.0484	0.8462	0.0061	1.1187	0.0050	0.9256	0.0117	0.7968	0.0358	0.9108
$\widetilde{F}_{DRs}$	0.0204	0.4797	0.0181	0.3530	0.0224	0.2721	0.0074	0.3427	0.0202	1.1169	0.0204	0.9244	0.0182	0.7953	0.0149	0.7064
$\widetilde{F}_{DRl}$	0.0103	0.4599	0.0066	0.3321	0.0170	0.2529	0.0190	0.3206	0.0082	1.0779	0.0098	0.8847	0.0113	0.7619	0.0119	0.6246
$\widetilde{F}_{DRsl}$	0.0149	0.4408	0.0124	0.3245	0.0182	0.2508	0.0096	0.2987	0.0135	1.0326	0.0144	0.8464	0.0133	0.7270	0.0106	0.6459
$\widetilde{F}_{DRls}$	0.0160	2.1553	0.0146	1.6542	0.0102	1.0080	0.0044	0.7626	0.0151	4.2860	0.0161	3.7163	0.0130	3.0589	0.0094	2.4749
$\widetilde{F}_{MCs}$	0.0148	0.4593	0.0113	0.3564	0.0161	0.3128	0.0123	0.5516	0.0056	1.0395	0.0088	0.8470	0.0122	0.7482	0.0146	0.7853
$\widetilde{F}_{MCl}$	0.0119	0.4712	0.0133	0.3682	0.0212	0.3237	0.0156	0.5806	0.0072	1.0603	0.0084	0.8729	0.0132	0.7729	0.0110	0.7760

Table 6 Populations generated from the models with not identically distributed error components. Simulated AVRB and AVRE in case of Poisson sampling with inclusion probabilities proportional to $\sqrt{x_i}$ and expected sample size $n = 50$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0639	1.0000	0.1218	1.0000	0.1803	1.0000	0.0458	1.0000	0.0890	1.0000	0.0807	1.0000	0.0742	1.0000	0.0628	1.0000
$\widehat{F}_{CD}$	0.2138	1.5061	0.2035	1.4102	0.1628	0.8630	0.1409	1.9677	0.1548	1.1235	0.1649	1.1759	0.1645	1.1959	0.0489	1.0121
$\widehat{F}_{DHs}$	0.2250	1.5222	0.2228	1.4813	0.1975	0.9768	0.0382	0.3215	0.1615	1.1427	0.1788	1.2183	0.1881	1.2724	0.0645	0.6159
$\widehat{F}_{DHl}$	0.2238	1.5244	0.2218	1.4973	0.1939	0.9595	0.1312	0.6522	0.1619	1.1294	0.1772	1.2036	0.1837	1.2502	0.0296	0.5524
$\widehat{F}_{Ks}$	0.0164	2.6265	0.0233	2.4452	0.0260	1.7486	0.0574	1.0619	0.0133	2.2697	0.0150	2.3448	0.0131	2.2212	0.0489	1.8287
$\widehat{F}_{Kl}$	0.0097	1.0370	0.0201	0.9163	0.0326	0.7124	0.1249	1.4799	0.0151	0.9744	0.0134	1.0067	0.0150	1.0201	0.0900	1.1127
$\widehat{F}_{CDWs}$	0.0131	2.6273	0.0178	2.4366	0.0247	1.7440	0.0552	1.0594	0.0126	2.2721	0.0141	2.3463	0.0120	2.2231	0.0474	1.8302
$\widehat{F}_{CDWl}$	0.0184	1.0394	0.0224	0.9126	0.0366	0.7091	0.1191	1.5369	0.0147	0.9801	0.0129	1.0113	0.0146	1.0257	0.0863	1.1259
$\widehat{F}_{DRs}$	0.0465	0.8911	0.0372	0.7172	0.0263	0.5223	0.0138	0.3620	0.0354	0.8883	0.0387	0.9003	0.0431	0.9108	0.0222	0.6618
$\widehat{F}_{DRl}$	0.0420	0.8356	0.0415	0.6939	0.0265	0.5074	0.0550	0.3864	0.0313	0.8511	0.0315	0.8571	0.0357	0.8677	0.0204	0.5819
$\widehat{F}_{DRsl}$	0.0563	0.8073	0.0519	0.6522	0.0399	0.4784	0.0295	0.3072	0.0480	0.8377	0.0487	0.8397	0.0535	0.8517	0.0231	0.5713
$\widehat{F}_{DRls}$	0.0234	2.4520	0.0243	2.1890	0.0186	1.8389	0.0268	0.5551	0.0187	1.9940	0.0187	2.0380	0.0188	2.0555	0.0154	1.3556
$\widetilde{F}_{CDs}$	0.0609	0.9602	0.1053	0.9209	0.1598	0.8903	0.0527	1.0135	0.0921	0.9901	0.0847	0.9903	0.0784	0.9875	0.0637	0.9928
$\widetilde{F}_{DHs}$	0.1076	1.0126	0.0761	0.7619	0.0435	0.5079	0.0188	0.4959	0.1133	0.9977	0.1166	1.0266	0.1158	1.0336	0.0548	0.7415
$\widetilde{F}_{DHl}$	0.1151	1.0265	0.0824	0.7712	0.0452	0.5002	0.0286	0.5423	0.1180	0.9995	0.1219	1.0292	0.1205	1.0337	0.0532	0.7328
$\widetilde{F}_{Ks}$	0.0082	2.5605	0.0074	2.3708	0.0172	1.6940	0.0361	0.9470	0.0068	2.2214	0.0066	2.2928	0.0060	2.1654	0.0185	1.7429
$\widetilde{F}_{Kl}$	0.0063	0.9505	0.0116	0.8207	0.0262	0.6394	0.0411	0.7973	0.0040	0.9056	0.0043	0.9333	0.0073	0.9449	0.0204	0.8395
$\widetilde{F}_{CDWs}$	0.0082	2.5599	0.0075	2.3700	0.0172	1.6933	0.0361	0.9469	0.0068	2.2188	0.0066	2.2925	0.0059	2.1656	0.0185	1.7421
$\widetilde{F}_{CDWl}$	0.0063	0.9493	0.0116	0.8196	0.0261	0.6380	0.0411	0.7970	0.0040	0.9051	0.0042	0.9330	0.0072	0.9444	0.0205	0.8390
$\widetilde{F}_{DRs}$	0.0275	0.9691	0.0188	0.7946	0.0110	0.5843	0.0102	0.4903	0.0266	0.9358	0.0287	0.9614	0.0282	0.9693	0.0173	0.7684
$\widetilde{F}_{DRl}$	0.0089	0.9281	0.0080	0.7852	0.0166	0.5916	0.0188	0.4637	0.0072	0.8915	0.0087	0.9141	0.0086	0.9235	0.0096	0.7153
$\widetilde{F}_{DRsl}$	0.0193	0.8952	0.0136	0.7349	0.0120	0.5438	0.0108	0.4547	0.0168	0.8721	0.0192	0.8941	0.0194	0.9022	0.0108	0.7023
$\widetilde{F}_{DRls}$	0.0122	2.5050	0.0110	2.3019	0.0032	1.8222	0.0110	0.6952	0.0076	1.9845	0.0094	2.1015	0.0091	2.2049	0.0043	1.4637
$\widetilde{F}_{MCs}$	0.0032	0.8863	0.0065	0.7472	0.0123	0.6265	0.0117	0.6203	0.0045	0.8848	0.0052	0.8958	0.0047	0.9090	0.0115	0.7971
$\widetilde{F}_{MCl}$	0.0034	0.9168	0.0093	0.7927	0.0198	0.6628	0.0102	0.6447	0.0069	0.8991	0.0072	0.9148	0.0042	0.9193	0.0086	0.8198

Table 7 Populations generated from the models with identically distributed error components. Simulated AVRB and AVRE in case of Poisson sampling with inclusion probabilities proportional to $\sqrt{x_i}$ and expected sample size $n = 100$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.2216	1.0000	0.2583	1.0000	0.2808	1.0000	0.0364	1.0000	0.0892	1.0000	0.1268	1.0000	0.1541	1.0000	0.0248	1.0000
$\widehat{F}_{CD}$	0.0166	0.1265	0.0451	0.1000	0.0910	0.1477	0.1650	4.9404	0.0219	0.5447	0.0297	0.3835	0.0416	0.3217	0.1476	3.4673
$\widehat{F}_{DHs}$	0.0296	0.1896	0.0207	0.1305	0.0213	0.1009	0.0188	0.1711	0.0194	0.6591	0.0210	0.4711	0.0233	0.3788	0.0188	0.4407
$\widehat{F}_{DHl}$	0.0232	0.1732	0.0156	0.1150	0.0218	0.0887	0.0973	0.7211	0.0184	0.6186	0.0204	0.4355	0.0207	0.3443	0.0680	0.5362
$\widehat{F}_{Ks}$	0.0182	4.7892	0.0194	3.5818	0.0135	1.5051	0.0407	1.8477	0.0194	10.4194	0.0174	7.9926	0.0106	5.2169	0.0391	2.7965
$\widehat{F}_{Kl}$	0.0265	0.3459	0.0390	0.2592	0.0405	0.1989	0.0938	1.4383	0.0101	1.0060	0.0136	0.7516	0.0129	0.6004	0.0953	1.5121
$\widehat{F}_{CDWs}$	0.0206	4.7950	0.0164	3.5564	0.0083	1.5029	0.0387	1.8394	0.0193	10.4226	0.0182	7.9944	0.0113	5.2189	0.0373	2.7918
$\widehat{F}_{CDWl}$	0.0074	0.3274	0.0139	0.2364	0.0210	0.1789	0.0884	1.4072	0.0083	1.0060	0.0095	0.7492	0.0080	0.5991	0.0902	1.5004
$\widehat{F}_{DRs}$	0.0128	0.2813	0.0126	0.1895	0.0169	0.1439	0.0081	0.2384	0.0091	0.8895	0.0077	0.6647	0.0092	0.5344	0.0105	0.5536
$\widehat{F}_{DRl}$	0.0165	0.2423	0.0094	0.1539	0.0175	0.1111	0.0179	0.1722	0.0103	0.8344	0.0106	0.6071	0.0104	0.4766	0.0198	0.4232
$\widehat{F}_{DRsl}$	0.0141	0.2331	0.0123	0.1558	0.0188	0.1176	0.0096	0.1632	0.0128	0.7766	0.0104	0.5691	0.0113	0.4558	0.0148	0.4509
$\widehat{F}_{DRls}$	0.0127	0.4382	0.0126	0.3050	0.0166	0.1991	0.0120	0.6487	0.0102	1.6703	0.0088	1.0936	0.0089	0.8323	0.0109	0.9021
$\widetilde{F}_{CDs}$	0.1924	0.7621	0.2256	0.7379	0.2410	0.6803	0.0616	1.2041	0.0752	0.8690	0.1070	0.8203	0.1281	0.7680	0.0463	1.1012
$\widetilde{F}_{DHs}$	0.0132	0.2556	0.0068	0.1835	0.0121	0.1501	0.0069	0.2820	0.0146	0.7963	0.0145	0.5867	0.0127	0.4744	0.0083	0.6155
$\widetilde{F}_{DHl}$	0.0084	0.2435	0.0033	0.1744	0.0142	0.1430	0.0149	0.3073	0.0083	0.7534	0.0079	0.5509	0.0066	0.4442	0.0094	0.5663
$\widetilde{F}_{Ks}$	0.0195	4.7631	0.0152	3.5616	0.0043	1.4937	0.0259	1.7240	0.0186	10.3618	0.0173	7.9508	0.0102	5.1862	0.0117	2.6740
$\widetilde{F}_{Kl}$	0.0040	0.3051	0.0057	0.2175	0.0129	0.1662	0.0281	0.6812	0.0042	0.9335	0.0042	0.6930	0.0056	0.5560	0.0202	0.8085
$\widetilde{F}_{CDWs}$	0.0195	4.7630	0.0151	3.5614	0.0043	1.4934	0.0260	1.7238	0.0186	10.3610	0.0172	7.9511	0.0102	5.1859	0.0117	2.6737
$\widetilde{F}_{CDWl}$	0.0041	0.3048	0.0057	0.2172	0.0130	0.1659	0.0281	0.6807	0.0042	0.9330	0.0042	0.6928	0.0055	0.5555	0.0202	0.8088
$\widetilde{F}_{DRs}$	0.0098	0.3131	0.0066	0.2189	0.0125	0.1684	0.0032	0.3511	0.0091	0.9665	0.0089	0.7188	0.0084	0.5807	0.0069	0.7149
$\widetilde{F}_{DRl}$	0.0058	0.2903	0.0034	0.2017	0.0085	0.1491	0.0057	0.2876	0.0057	0.9111	0.0054	0.6731	0.0049	0.5362	0.0055	0.6172
$\widetilde{F}_{DRsl}$	0.0060	0.2827	0.0034	0.2009	0.0095	0.1527	0.0047	0.2969	0.0059	0.8850	0.0055	0.6535	0.0053	0.5260	0.0047	0.6469
$\widetilde{F}_{DRls}$	0.0082	0.4697	0.0075	0.3310	0.0119	0.2264	0.0040	0.6965	0.0089	1.9396	0.0081	1.1419	0.0094	0.8687	0.0065	1.1631
$\widetilde{F}_{MCs}$	0.0090	0.2906	0.0083	0.2176	0.0112	0.1892	0.0092	0.5570	0.0038	0.8799	0.0057	0.6519	0.0084	0.5425	0.0091	0.7731
$\widetilde{F}_{MCl}$	0.0059	0.2857	0.0072	0.2145	0.0116	0.1880	0.0093	0.5446	0.0025	0.8752	0.0036	0.6447	0.0071	0.5360	0.0078	0.7423

Table 8 Populations generated from the models with not identically distributed error components. Simulated AVRB and AVRE in case of Poisson sampling with inclusion probabilities proportional to $\sqrt{x_i}$ and expected sample size $n = 100$.

	$\sigma = 0.1$								$\sigma = 0.3$							
Model	M1		M2		M3		M4		M1		M2		M3		M4	
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
Estimator																
$\widehat{F}_{HT}$	0.0667	1.0000	0.1205	1.0000	0.1801	1.0000	0.0456	1.0000	0.0935	1.0000	0.0843	1.0000	0.0775	1.0000	0.0648	1.0000
$\widehat{F}_{CD}$	0.2117	2.1113	0.2009	1.8571	0.1594	0.7942	0.1376	3.0157	0.1501	1.2175	0.1602	1.3119	0.1596	1.3582	0.0452	1.1993
$\widehat{F}_{DHs}$	0.2234	2.2153	0.2213	2.0952	0.1962	1.0038	0.0223	0.3238	0.1608	1.3066	0.1774	1.4495	0.1860	1.5611	0.0706	0.6962
$\widehat{F}_{DHl}$	0.2175	2.1195	0.2165	2.0101	0.1903	0.9412	0.0762	0.4957	0.1557	1.2283	0.1717	1.3675	0.1791	1.4718	0.0498	0.6079
$\widehat{F}_{Ks}$	0.0095	7.3445	0.0159	4.5870	0.0178	1.5644	0.0397	1.9689	0.0102	3.9851	0.0113	4.2563	0.0094	5.2752	0.0311	3.3784
$\widehat{F}_{Kl}$	0.0113	0.8719	0.0229	0.6788	0.0323	0.4401	0.0968	1.2964	0.0119	0.8148	0.0123	0.8470	0.0120	0.8653	0.0755	1.0272
$\widehat{F}_{CDWs}$	0.0088	7.3453	0.0160	4.5926	0.0172	1.5651	0.0383	1.9624	0.0099	3.9846	0.0108	4.2569	0.0091	5.2760	0.0303	3.3746
$\widehat{F}_{CDWl}$	0.0080	0.8767	0.0179	0.6759	0.0300	0.4380	0.0925	1.2849	0.0111	0.8176	0.0111	0.8509	0.0108	0.8702	0.0730	1.0343
$\widehat{F}_{DRs}$	0.0254	0.7847	0.0186	0.5691	0.0137	0.3554	0.0137	0.3497	0.0229	0.7887	0.0217	0.8000	0.0251	0.8167	0.0158	0.6162
$\widehat{F}_{DRl}$	0.0275	0.7206	0.0239	0.5161	0.0165	0.3162	0.0282	0.2868	0.0240	0.7523	0.0238	0.7557	0.0265	0.7694	0.0186	0.5002
$\widehat{F}_{DRsl}$	0.0390	0.7201	0.0335	0.5141	0.0266	0.3227	0.0195	0.2968	0.0359	0.7546	0.0355	0.7567	0.0390	0.7759	0.0205	0.5327
$\widehat{F}_{DRls}$	0.0149	2.8065	0.0071	1.1641	0.0121	0.5696	0.0158	0.6482	0.0163	2.3724	0.0169	2.4223	0.0182	2.6667	0.0165	1.3177
$\widetilde{F}_{CDs}$	0.0646	0.9128	0.0865	0.7928	0.1392	0.7337	0.0582	1.1168	0.0992	0.9966	0.0916	0.9959	0.0858	0.9927	0.0672	1.0204
$\widetilde{F}_{DHs}$	0.1153	1.1837	0.0828	0.7661	0.0495	0.3767	0.0191	0.4793	0.1198	1.0818	0.1233	1.1404	0.1237	1.1703	0.0609	0.7605
$\widetilde{F}_{DHl}$	0.1209	1.2026	0.0883	0.7705	0.0527	0.3660	0.0209	0.4789	0.1223	1.0652	0.1264	1.1286	0.1271	1.1612	0.0599	0.7344
$\widetilde{F}_{Ks}$	0.0065	7.3022	0.0110	4.5463	0.0120	1.5446	0.0192	1.8769	0.0056	3.9544	0.0071	4.2227	0.0046	5.2390	0.0090	3.3281
$\widetilde{F}_{Kl}$	0.0028	0.8192	0.0055	0.6177	0.0133	0.3968	0.0234	0.6523	0.0024	0.7729	0.0029	0.8009	0.0034	0.8173	0.0128	0.7070
$\widetilde{F}_{CDWs}$	0.0066	7.3018	0.0110	4.5465	0.0119	1.5434	0.0193	1.8758	0.0056	3.9539	0.0070	4.2221	0.0047	5.2385	0.0090	3.3262
$\widetilde{F}_{CDWl}$	0.0028	0.8183	0.0055	0.6170	0.0132	0.3960	0.0234	0.6525	0.0024	0.7725	0.0029	0.8008	0.0034	0.8168	0.0128	0.7069
$\widetilde{F}_{DRs}$	0.0127	0.8605	0.0071	0.6353	0.0096	0.3969	0.0051	0.4654	0.0119	0.8249	0.0133	0.8500	0.0129	0.8641	0.0090	0.7042
$\widetilde{F}_{DRl}$	0.0054	0.8068	0.0026	0.5998	0.0076	0.3733	0.0071	0.4121	0.0047	0.7763	0.0059	0.7992	0.0053	0.8161	0.0054	0.6258
$\widetilde{F}_{DRsl}$	0.0088	0.7997	0.0040	0.5887	0.0075	0.3693	0.0055	0.4282	0.0081	0.7700	0.0093	0.7926	0.0090	0.8068	0.0061	0.6429
$\widetilde{F}_{DRls}$	0.0075	3.4419	0.0046	1.3715	0.0101	0.6807	0.0031	0.9030	0.0074	2.5704	0.0086	2.5339	0.0079	2.9988	0.0070	1.2865
$\widetilde{F}_{MCs}$	0.0020	0.7695	0.0047	0.6018	0.0093	0.4271	0.0076	0.5996	0.0059	0.7682	0.0054	0.7768	0.0042	0.7996	0.0063	0.7288
$\widetilde{F}_{MCl}$	0.0014	0.7652	0.0058	0.5954	0.0116	0.4251	0.0060	0.5883	0.0063	0.7510	0.0056	0.7699	0.0036	0.7916	0.0047	0.7036

References

1. Cassel, C. M., Saerndal, C. E., and Wretman, J. H. (1976). Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations. *Biometrika*, 63(3), 615–620.
2. Chambers, R.L., Dorfman, A.H., Hall, P. (1992). Properties of Estimators of the Finite Population Distribution Function. *Biometrika*, 79(3), 577–582
3. Chambers, R.L., Dorfman, A.H., Wehrly, T. (1993). Bias robust estimation in finite populations using non-parametric calibration. *J. Amer. Statist. Assoc.*, 88(421), 268–277
4. Chambers, R.L., Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73(3), 597–604
5. Chen, J., Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Stat. Sin.*, 12, 1223–1239
6. Deville, J.C., Saerndal C.E. (1992). Calibration estimators in survey sampling. *J. Am. Stat. Assoc.*, 87, 376–382
7. Dorfman, A.H., Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *The Annals of Statistics*, 21(3), 1452–1475.
8. Johnson, A.A., Breidt, F.J., Opsomer, J.D. (2008). Estimating Distribution Functions from Survey Data Using Nonparametric Regression. *J. Stat. Theory Pract.*, 2(3), 419–431.
9. Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. In *Proceedings of the section on survey research methods* (pp. 280-285), Amer. Statist. Assoc., Alexandria, VA.
10. Montanari, G.E., Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *J. Am. Stat. Assoc.*, 100, 1429–1442
11. Montanari, G.E., Ranalli, M.G. (2005). Nonparametric methods for sample surveys of environmental populations. Available at http://old.sis-statistica.org/files/pdf/atti/CIMe0905p147-158.pdf
12. Rao, J.N.K., Kovar J.G., Mantel H.J. (1990). On Estimating Distribution Functions and Quantiles from Survey Data Using Auxiliary Information. *Biometrika*, 77(2), 365–375.
13. Rueda, M., Martinez, S., Martinez, H., Arcos, A. (2007). Estimation of the distribution function with calibration methods. *J. Stat. Plan. Inference*, 137(2), 435–448
14. Rueda, M., Snchez-Borrego, I., Arcos, A., Martinez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika*, 71, 33–44
15. Saerndal, C. E. (1980). On π-inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. *Biometrika*, 67(3), 639–650.
16. Saerndal C.E. (2007). The calibration approach in survey theory and practice. *Surv. Methodol.*, 33, 99–119
17. Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90, 937–951
18. Wu, C., Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *J. Am. Stat. Assoc.*, 96, 185–193