Nonlinear Analysis: Modelling and Control, 2009, Vol. 14, No. 3, 315–331

Estimation of a Distribution Function under Sampling on Two Occasions

V. Chadyšas

Vilnius Gediminas Technical University
Saulėtekio ave. 11, LT-10223 Vilnius, Lithuania
viktoras.chadysas@fm.vgtu.lt

Received: 2009-01-19 **Revised:** 2009-05-20 **Published online:** 2009-09-11

Abstract. Estimation of the distribution function under sampling on two occasions with a simple random sampling design on each occasion is investigated. Composite regression and ratio type estimators are considered, using values of the study variable as auxiliary information obtained on the first occasion. The optimal estimator, in the sense of minimal variance, is also obtained. A simulation study, based on the real population data, is performed and the proposed estimators are compared by a simple estimator for a distribution function.

Keywords: distribution function, sampling on two occasions, auxiliary information, regression estimator, ratio estimator.

1 Introduction

Consider a finite population $\mathcal{U} = \{1, \ldots, N\}$. Let y be the study variable, defined on the population $\mathcal{U}$ and taking values $\{y_1, \ldots, y_N\}$. The values of the variable y are not known. We are interested in the estimation of the finite population distribution function of the study variable y

$$F_y(z) = \frac{\sharp A_z}{N},$$

where for any given number z $(-\infty < z < \infty)$, the set $A_z = \{l \in \mathcal{U} : \ y_l \leq z\}$, and $\sharp A_z$ denotes the number of elements in the set A_z (see [1,2]). Such a function $F_y(z)$ may be of considerable interest when y is a measure of income and the population units are individuals or households.

In sample surveys, supplementary information is often used in the estimation stage to increase the precision of estimators of the population mean or total. Since $F(z)$ is simply a population proportion for any given value of z, usual methods for estimating the means such as the ratio and regression estimators taking advantage of auxiliary information can be used. Recently, several estimators of the population distribution function have been

proposed, using auxiliary information in the estimation stage (see [3–7]). Most of the studies related to a distribution function have been developed by assuming simple random sampling or a stratified simple random sampling design.

When the investigation deals with variables such as income, sometimes the same population is sampled repeatedly on several occasions and the same study variable is measured on each occasion. Repeated sampling of population is a quite common sampling procedure in the official statistics.

Cochran (see [8, chapter 12]) considered sampling on two occasions, using random sampling at each of the occasions. He has found that current estimates might be improved by using the first occasion data. Some problems of estimator construction for sampling on two occasions have been discussed (see [8–10]). In all the studies, the parameter estimated is a mean.

In this paper, we investigate sampling on two occasions, concentrating on the estimators of the distribution function. The aim of this paper is, first, to obtain some estimators of the distribution function under sampling on two occasions: the ratio and regression estimators; second, to obtain optimal composite estimators in the sense of minimizing the variance of the estimators; third, to investigate how the sample matching fraction influences precision of the distribution function estimates using a sampling scheme on two occasions, and, finally, to illustrate the theoretical results by simulation study.

2 Estimation of the distribution function using a scheme of two occasions

Suppose we have a finite population $\mathcal{U} = \{1, \ldots, N\}$ of size N, which is assumed to retain its composition over two-time periods.

Let us denote the study variable on the second occasion by y, and the same variable on the first occasion by x with the values y_i, and x_i. Denote by n' the sample size on the first occasion.

On the second occasion, two independent samples are drawn, one being matched with the sample of the first occasion and the other unmatched. The matched sample is a subsample of size m, drawn from the previously selected n' units, and the unmatched sample of size u is drawn from $N - n'$ remaining units. Thus, the total sample on the current occasion consists of $n = m + u$ units.

So, we have a two-phase sampling scheme. The first-phase sample s' of size n' is drawn according to a certain sampling design with $p(s')$, i.e., the probability of s' being chosen. The corresponding first and second order inclusion probabilities are π'_i, π'_{ij}, for $i, j \in \mathcal{U}$.

Given s', on the second occasion, a matched sample s_m of size m is drawn from s' according to a certain sampling design, such that $p(s_m|s')$ is the conditional probability of choosing s_m. The corresponding first and second order inclusion probabilities are $\pi_{i|s'}$, $\pi_{ij|s'}$.

The unmatched sample s_u of size u is drawn from $\mathcal{U} \setminus s' = s'^c$ in accordance with a certain sampling design, such that $p(s_u|s'^c)$ is the conditional probability of choosing s_u.

The corresponding first and second order inclusion probabilities are $\pi_{i|s'^c}$, $\pi_{ij|s'^c}$. The whole sample on the current occasion is $s = s_m \cup s_u$.

We are interested in estimation of the finite population distribution function using a two occasion scheme, when a simple random sampling design is used at each of the occasions.

2.1 Simple estimator

Let us define an indicator variable $h(z)$ with the values

$$h_i(z) = \begin{cases} 1, & \text{if } y_i \le z, \\ 0, & \text{if } y_i > z, \quad -\infty < z < \infty, \end{cases}$$

$i = 1, 2, \ldots, N$, and its total $t_{h(z)} = \sum_{i=1}^{N} h_i(z)$. Then the distribution function of the study variable y can be expressed as:

$$F_y(z) = \frac{t_{h(z)}}{N} = \frac{1}{N} \sum_{i=1}^{N} h_i(z). \tag{1}$$

The whole second phase sample s consists of two samples s_m and s_u, for sampling on two occasions each of them being a two-phase sample:

$$\mathcal{U} \to s' \to s_m,$$
$$\mathcal{U} \to \mathcal{U} \setminus s' = s'^c \to s_u.$$

Under two-phase sampling, Särndal et al. (see [2, chapter 9]) have shown, that the usual Horvitz-Thompson type estimator of the population total cannot always be used in practice, because the inclusion probabilities, associated with the second-phase sample, should be known for each first-phase sample. The use of π^* estimators is a possible alternative, proposed by Särndal et al. (see [2, chapter 9]), for the problem of estimation of the population total. Using this idea, Rueda et al. (see [11]) have presented the quantities

$$\begin{aligned} \pi_i^* &= \mathrm{P}(s' \colon i \in s')\mathrm{P}(s_m \colon i \in s_m \mid s') + \mathrm{P}(s'^c \colon i \in s'^c)\mathrm{P}(s_u \colon i \in s_u \mid s'^c) \\ &= \pi_i' \pi_{i|s'} + \pi_i'^c \pi_{i|s'^c}, \end{aligned} \tag{2}$$

where $\pi_i'^c = 1 - \pi_i'$.

Using the samples s_u and s_m, the following unbiased π^* estimator of the distribution function (1) can be constructed

$$\begin{aligned} \widehat{F}_y(z) &= \frac{1}{N} \sum_{i \in s} \frac{h_i(z)}{\pi_i^*} = \frac{1}{N} \sum_{i \in s_m} \frac{h_i(z)}{\pi_i^*} + \frac{1}{N} \sum_{i \in s_u} \frac{h_i(z)}{\pi_i^*} \\ &= \frac{1}{N} \sum_{i \in s_m} \frac{\pi_i' \pi_{i|s'}}{\pi_i^*} \frac{h_i(z)}{\pi_i' \pi_{i|s'}} + \frac{1}{N} \sum_{i \in s_u} \frac{\pi_i'^c \pi_{i|s'^c}}{\pi_i^*} \frac{h_i(z)}{\pi_i'^c \pi_{i|s'^c}}, \end{aligned} \tag{3}$$

for any sampling designs on both occasions.

Let us introduce new notation:

$$d_{1i} = \frac{\pi'_i \pi_{i|s'}}{\pi^*_i},\ i \in s_m, \quad \widehat{t}_{h(z)_m} = \sum_{i \in s_m} \frac{h_i(z)}{\pi'_i \pi_{i|s'}}, \quad \text{unbiased,}$$

$$d_{2i} = \frac{\pi'^c_i \pi_{i|s'^c}}{\pi^*_i},\ i \in s_u, \quad \widehat{t}_{h(z)_u} = \sum_{i \in s_u} \frac{h_i(z)}{\pi'^c_i \pi_{i|s'^c}}, \quad \text{unbiased.}$$

The coefficients d_{1i}, d_{2i} do not depend on i for design of a simple random sample on each occasions, for a two-occasion sampling scheme

$$d_{1i} = d_1,\ \ i \in s_m, \quad d_{2i} = d_2,\ \ i \in s_u.$$

Under the new notation, introduced before, the estimator of distribution function (3) can be expressed as

$$\widehat{F}_y(z) = \frac{1}{N} d_1 \widehat{t}_{h(z)_m} + \frac{1}{N} d_2 \widehat{t}_{h(z)_u}. \tag{4}$$

Assume that s' is a simple random sample from the population $\mathcal{U}$ and its complement s'^c is also a simple random sample from the population $\mathcal{U}$. s_m is a simple random sample from s' and s_u is a simple random sample from s'^c. Then the first and second stage inclusion probabilities are calculated as follows:

$$\pi'_i = \frac{n'}{N}, \quad \pi'_{ij} = \frac{n'}{N}\frac{n'-1}{N-1}, \quad \pi_{i|s'} = \frac{m}{n'}, \quad \pi_{ij|s'} = \frac{m(m-1)}{n'(n'-1)},$$

$$\pi'^c_i = \frac{N-n'}{N}, \quad \pi_{i|s'^c} = \frac{u}{N-n'}, \quad \pi_{ij|s'^c} = \frac{u(u-1)}{(N-n')(N-n'-1)},$$

$$\pi^*_i = \pi'_i \pi_{i|s'} + \pi'^c_i \pi_{i|s'^c} = \frac{n'}{N}\frac{m}{n'} + \frac{N-n'}{N}\frac{u}{N-n'} = \frac{m}{N} + \frac{u}{N} = \frac{n}{N}$$

and the coefficients d_1 and d_2 are:

$$d_1 = \frac{m}{n}, \quad d_2 = \frac{u}{n}.$$

In the case of simple random sampling, on each of the two occasions the estimator (4) of the distribution function can be rewritten as

$$\widehat{F}_y(z) = \frac{m}{n}\frac{1}{N}\widehat{t}_{h(z)_m} + \frac{u}{n}\frac{1}{N}\widehat{t}_{h(z)_u} = \frac{m}{n}\overline{h(z)}_m + \frac{u}{n}\overline{h(z)}_u, \tag{5}$$

where

$$\overline{h(z)}_m = \frac{1}{m}\sum_{i \in s_m} h_i(z), \quad \overline{h(z)}_u = \frac{1}{u}\sum_{i \in s_u} h_i(z).$$

In the case of simple random sampling, on each of the two occasions, the resulting sample of size $n = m + u$ is also simple random sample. The variance $\mathrm{Var}(\widehat{F}_y(z))$ of the distribution function $F_y(z)$ estimator $\widehat{F}_y(z)$ (5) is expressed:

$$\mathrm{Var}\big(\widehat{F}_y(z)\big) = \left(1 - \frac{n}{N}\right)\frac{s^2_{h(z)}}{n}, \tag{6}$$

where

$$s^2_{h(z)} = \frac{1}{N-1}\sum_{i=1}^{N}\big(h_i(z) - \mu_{h(z)}\big)^2, \quad \mu_{h(z)} = \frac{1}{N}\sum_{i=1}^{N} h_i(z).$$

Remark 1. We use the unbiased variance estimator $\widehat{\mathrm{Var}}(\widehat{F}_y(z))$ of the distribution function estimator (5) by replacing $s^2_{h(z)}$ in variance expression (6) with

$$\widehat{s}^2_{h(z)_n} = \frac{1}{n-1}\sum_{i\in s}\big(h_i(z) - \overline{h(z)}_n\big)^2, \quad \overline{h(z)}_n = \frac{1}{n}\sum_{i\in s} h_i(z).$$

2.2 Regression type estimator

In sample surveys, auxiliary information is often used at the estimation stage to increase the accuracy of estimators. Using sampling on two occasions we can construct distribution function estimators with x_i values from the first occasion sample as auxiliary information.

Let us define a new indicator variable $g(z)$ with the values

$$g_i(z) = \begin{cases} 1, & \text{if } x_i \le z, \\ 0, & \text{if } x_i > z, \end{cases}$$

$i = 1, 2, \ldots, N$, and the total $t_{g(z)} = \sum_{i=1}^{N} g_i(z)$. Then the distribution function $F_x(z)$ can be expressed as:

$$F_x(z) = \frac{t_{g(z)}}{N} = \frac{1}{N}\sum_{i=1}^{N} g_i(z). \tag{7}$$

Using the first occasion sample s' and the matched sample s_m, we can form a regression type estimator of the distribution function

$$\widehat{F}^{reg}_{ym}(z) = \frac{1}{N}\widehat{t}^{reg}_{h(z)_m} = \frac{1}{N}\widehat{t}_{h(z)_m} + \frac{1}{N}b\big(\widehat{t}_{g(z)_{n'}} - \widehat{t}_{g(z)_m}\big), \tag{8}$$

with

$$\widehat{t}_{h(z)_m} = \sum_{i\in s_m}\frac{h_i(z)}{\pi_i\pi_{i|s'}}, \quad \widehat{t}_{g(z)_m} = \sum_{i\in s_m}\frac{g_i(z)}{\pi_i\pi_{i|s'}}, \quad \widehat{t}_{g(z)_{n'}} = \sum_{i\in s'}\frac{g_i(z)}{\pi_i}.$$

and b is some constant.

A second estimator $\widehat{F}_{yu}(z)$ of the distribution function $F_y(z)$ can be obtained from the unmatched sample s_u. It was already introduced in (5).

By a linear combination of $\widehat{F}^{reg}_{ym}(z)$ and $\widehat{F}_{yu}(z)$ we obtain a new type of composite regression estimator

$$\widehat{F}^{reg}_{y}(z) = \omega \frac{1}{N} \widehat{t}^{reg}_{h(z)_m} + (1-\omega)\frac{1}{N}\widehat{t}_{h(z)_u}, \tag{9}$$

where ω is a constant $(0 < \omega < 1)$. The variance of the term $\widehat{t}^{reg}_{h(z)_m}$ depends on the constant b. We can find b_{opt} by minimizing the variance $\mathrm{Var}(\widehat{t}^{reg}_{h(z)_m})$.

$$b_{opt} = \frac{s_{h(z)g(z)}}{s^2_{g(z)}} = \frac{\sum_{i=1}^{N}(h_i(z) - \mu_{h(z)})(g_i(z) - \mu_{g(z)})}{\sum_{i=1}^{N}(g_i(z) - \mu_{g(z)})^2}, \tag{10}$$

where

$$\mu_{h(z)} = \frac{1}{N}\sum_{i=1}^{N} h_i(z), \quad \mu_{g(z)} = \frac{1}{N}\sum_{i=1}^{N} g_i(z).$$

Since the values of indicator variables $h(z)$ and $g(z)$ are not known in the population as usual, we cannot calculate the coefficient b_{opt}, so we need to estimate it from a sample. The coefficient b_{opt} can be estimated by

$$\widehat{b}_{opt} = \frac{\widehat{s}_{h(z)g(z)}}{\widehat{s}^2_{g(z)}} = \frac{\sum_{i\in s_m}(h_i(z) - \overline{h(z)}_m)(g_i(z) - \overline{g(z)}_m)}{\sum_{i\in s_m}(g_i(z) - \overline{g(z)}_m)^2}, \tag{11}$$

where

$$\overline{g(z)}_m = \frac{1}{m}\sum_{i\in s_m} g_i(z),$$

$\overline{h(z)}_m$ has been defined in (5).

In the case of simple random sampling on each of two occasions, estimator (9) of the distribution function $F_y(z)$, using a two-occasion scheme, can be expressed:

$$\widehat{F}^{reg}_{y}(z) = \omega\Big(\overline{h(z)}_m + \widehat{b}_{opt}\big(\overline{g(z)}_{n'} - \overline{g(z)}_m\big)\Big) + (1-\omega)\overline{h(z)}_u, \tag{12}$$

where

$$\overline{g(z)}_{n'} = \frac{1}{n'}\sum_{i\in s'} g_i(z), \quad \overline{g(z)}_m = \frac{1}{m}\sum_{i\in s_m} g_i(z),$$

$\overline{h(z)}_m$ and $\overline{h(z)}_u$ have been defined earlier by equalities (5).

Proposition 1. *In the case of simple random sampling on each of the two occasions, an approximate variance* $\mathrm{AVar}(\widehat{F}_y^{reg}(z))$ *of regression type estimator* $\widehat{F}_y^{reg}(z)$ (12) *of the distribution function* $F_y(z)$ *is expressed:*

$$\mathrm{AVar}\big(\widehat{F}_y^{reg}(z)\big) = \omega^2 \frac{1}{N^2}\mathrm{AVar}\big(\widehat{t}^{reg}_{h(z)_m}\big) + (1-\omega)^2 \frac{1}{N^2}\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) + 2\omega(1-\omega)\frac{1}{N^2}\mathrm{Cov}\big(\widehat{t}^{reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big), \tag{13}$$

$$\mathrm{AVar}\big(\widehat{t}^{reg}_{h(z)_m}\big) = N^2\left(\left(1-\frac{n'}{N}\right)\frac{s^2_{h(z)}}{n'} + \left(1-\frac{m}{n'}\right)\frac{s^2_{D(z)}}{m}\right),$$

$$\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) = N^2\left(1-\frac{u}{N}\right)\frac{s^2_{h(z)}}{u}, \quad \mathrm{Cov}\big(\widehat{t}^{reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big) = -N s^2_{h(z)},$$

$$s^2_{D(z)} = \frac{1}{N-1}\sum_{i=1}^{N}\big(D_i(z) - \mu_{D(z)}\big)^2, \quad \mu_{D(z)} = \frac{1}{N}\sum_{i=1}^{N} D_i(z),$$

$s^2_{h(z)}$ *has been defined earlier in* (6) *and* $D_i(z) = h_i(z) - b_{opt}g_i(z)$.

Proof. The variance of the composite regression type estimator (9) equals

$$\mathrm{Var}\big(\widehat{F}_y^{reg}(z)\big) = \omega^2 \frac{1}{N^2}\mathrm{Var}\big(\widehat{t}^{reg}_{h(z)_m}\big) + (1-\omega)^2 \frac{1}{N^2}\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) + 2\omega(1-\omega)\frac{1}{N^2}\mathrm{Cov}\big(\widehat{t}^{reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big). \tag{14}$$

The known approximation of the $\mathrm{Var}(\widehat{t}^{reg}_{h(z)_m})$ (see [2]) is

$$\mathrm{AVar}\big(\widehat{t}^{reg}_{h(z)_m}\big) = \sum_{i,j\in\mathcal{U}}\big(\pi'_{ij} - \pi'_i\pi'_j\big)\frac{h_i(z)}{\pi'_i}\frac{h_j(z)}{\pi'_j} + \mathrm{E}\left(\sum_{i,j\in s'}\big(\pi_{ij|s'} - \pi_{i|s'}\pi_{j|s'}\big)\frac{D_i(z)}{\pi'_i\pi_{i|s'}}\frac{D_j(z)}{\pi'_j\pi_{j|s'}}\right), \tag{15}$$

with $D_i(z) = h_i(z) - b_{opt}g_i(z)$. $\mathrm{Var}(\widehat{t}_{h(z)_u})$ and covariance $\mathrm{Cov}(\widehat{t}^{reg}_{h(z)_m}, \widehat{t}_{h(z)_u}) = \mathrm{Cov}(\widehat{t}_{h(z)_m}, \widehat{t}_{h(z)_u})$ are expressed, respectively, as

$$\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) = \sum_{i,j\in\mathcal{U}}\big(\pi'^c_{ij} - \pi'^c_i\pi'^c_j\big)\frac{h_i(z)}{\pi'^c_i}\frac{h_j(z)}{\pi'^c_j} + \mathrm{E}\left(\sum_{i,j\in s'^c}\big(\pi_{ij|s'^c} - \pi_{i|s'^c}\pi_{j|s'^c}\big)\frac{h_i(z)}{\pi'^c_i\pi_{i|s'^c}}\frac{h_j(z)}{\pi'^c_j\pi_{j|s'^c}}\right) \tag{16}$$

and

$$\mathrm{Cov}\big(\widehat{t}^{\,reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big) = -\sum_{i,j\in\mathcal{U}} \big(\pi'_{ij} - \pi'_i\pi'_j\big)\frac{h_i(z)}{\pi'_i}\frac{h_j(z)}{\pi'^{c}_j}. \tag{17}$$

Replacing π values in (14), (15) by the corresponding values, obtained for a simple random sampling design on each of the two occasions, we obtain an expression of approximate variance (13) of the distribution function estimator (12). □

Remark 2. We use variance estimator $\widehat{\mathrm{Var}}(\widehat{F}^{reg}_y(z))$ of the composite regression type distribution function estimator (12), replacing $s^2_{h(z)}$ and $s^2_{D(z)}$ in the $\mathrm{AVar}(\widehat{t}^{\,reg}_{h(z)_m})$ of (13) by the estimates below

$$\widehat{s}^2_{h(z)_m} = \frac{1}{m-1}\sum_{i\in s_m}\big(h_i(z) - \overline{h(z)}_m\big)^2, \quad \overline{h(z)}_m = \frac{1}{m}\sum_{i\in s_m} h_i(z),$$

and

$$\widehat{s}^2_{D(z)_m} = \frac{1}{m-1}\sum_{i\in s_m}\big(\widehat{D}_i(z) - \overline{\widehat{D}(z)}_m\big)^2, \quad \overline{\widehat{D}(z)}_m = \frac{1}{m}\sum_{i\in s_m} \widehat{D}_i(z),$$

where $\widehat{D}_i(z) = h_i(z) - \widehat{b}_{opt}g_i(z)$.

In the $\mathrm{Var}(\widehat{t}_{h(z)_u})$, $s^2_{h(z)}$ is replaced by

$$\widehat{s}^2_{h(z)_u} = \frac{1}{u-1}\sum_{i\in s_u}\big(h_i(z) - \overline{h(z)}_u\big)^2, \quad \overline{h(z)}_u = \frac{1}{u}\sum_{i\in s_u} h_i(z),$$

and in the $\mathrm{Cov}(\widehat{t}^{\,reg}_{h(z)_m}, \widehat{t}_{h(z)_u})$, $s^2_{h(z)}$ is replaced by

$$\widehat{s}^2_{h(z)_n} = \frac{1}{n-1}\sum_{i\in s}\big(h_i(z) - \overline{h(z)}_n\big)^2, \quad \overline{h(z)}_n = \frac{1}{n}\sum_{i\in s} h_i(z).$$

We use a constant ω, in the expression of regression type estimator (12) of the distribution function $F_y(z)$. Its optimal value ω_{opt} can be found in the sense of minimal variance (13).

Proposition 2. *In the case of simple random sampling on each of the two occasions, the optimal value ω_{opt} in* (12) *is expressed:*

$$\omega_{opt} = \frac{\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) - \mathrm{Cov}\big(\widehat{t}^{\,reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big)}{\mathrm{Var}\big(\widehat{t}^{\,reg}_{h(z)_m}\big) + \mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) - 2\mathrm{Cov}\big(\widehat{t}^{\,reg}_{h(z)_m}, \widehat{t}_{h(z)_u}\big)} \tag{18}$$

on the two-occasion sampling scheme.

Proof. Differentiating $\mathrm{Var}(F_y^{reg}(z)$ in (14) with respect to the coefficient ω and equating the derivative to zero, we get the optimal value ω_{opt} of the coefficient ω. □

Replacing the coefficient ω by the coefficient ω_{opt} in the distribution function estimator $\widehat{F}_y^{reg}(z)$ given by (12), we obtain an optimal composite regression type estimator of the distribution function. In the case of simple random sampling on each of the two occasions:

$$\widehat{F}_{y\ opt}^{reg}(z) = \omega_{opt}\Big(\overline{h(z)}_m + \widehat{b}_{opt}\big(\overline{g(z)}_{n'} - \overline{g(z)}_m\big)\Big) + (1-\omega_{opt})\overline{h(z)}_u. \tag{19}$$

Proposition 3. *In the case of simple random sampling on each of the two occasions, the approximate minimal variance* $\mathrm{AVar}(\widehat{F}_{y\ opt}^{reg}(z))_{min}$ *of the regression type estimator* $\widehat{F}_{y\ opt}^{reg}(z)$ (19) *of the distribution function* $F_y(z)$ *is expressed:*

$$\mathrm{AVar}\big(\widehat{F}_{y\ opt}^{reg}(z)\big)_{min} = \frac{1}{N^2}\left(\frac{\mathrm{Var}_1\mathrm{Var}_2 - \mathrm{Cov}^2}{\mathrm{Var}_1 + \mathrm{Var}_2 - 2\mathrm{Cov}}\right), \tag{20}$$

where $\mathrm{Var}_1 = \mathrm{AVar}(\widehat{t}_{h(z)_m}^{reg})$, $\mathrm{Var}_2 = \mathrm{Var}(\widehat{t}_{h(z)_u})$, $\mathrm{Cov} = \mathrm{Cov}(\widehat{t}_{h(z)_m}^{reg}, \widehat{t}_{h(z)_u})$.

Proof. By inserting the optimal value ω_{opt} (18) of ω into the expression of approximate variance (13), we obtain (20). □

Remark 3. The coefficient w_{opt} depends on unknown variances and the covariance, and we estimate it by

$$\widehat{\omega}_{opt} = \frac{\widehat{\mathrm{Var}}(\widehat{t}_{h(z)_u}) - \widehat{\mathrm{Cov}}(\widehat{t}_{h(z)_m}^{reg}, \widehat{t}_{h(z)_u})}{\widehat{\mathrm{Var}}(\widehat{t}_{h(z)_m}^{reg}) + \widehat{\mathrm{Var}}(\widehat{t}_{h(z)_u}) - 2\widehat{\mathrm{Cov}}(\widehat{t}_{h(z)_m}^{reg}, \widehat{t}_{h(z)_u})}. \tag{21}$$

We use the approximate minimal variance estimator $\widehat{\mathrm{Var}}(\widehat{F}_y^{reg}(z))_{min}$ of the composite optimal regression type distribution function estimator (19) replacing Var_1, Var_2, and Cov in $\mathrm{AVar}(\widehat{t}_{h(z)_m}^{reg})_{min}$ of (20) by the corresponding estimators $\widehat{\mathrm{Var}}_1$, $\widehat{\mathrm{Var}}_2$, and $\widehat{\mathrm{Cov}}$.

2.3 Ratio type estimator

A particular case within the regression type estimator is the ratio type estimator. Distribution function estimators of the regression type and ratio type differ in the choice coefficient b in (8).

Using the first occasion sample s' and the matched sample s_m, we can form a ratio type estimator of the distribution function

$$\widehat{F}_{ym}^{r}(z) = \frac{1}{N}\widehat{t}_{h(z)_m}^{r} = \frac{1}{N}\widehat{t}_{g(z)_{n'}}\widehat{R}(z), \tag{22}$$

where

$$\widehat{t}_{g(z)_{n'}} = \sum_{i\in s'} \frac{g_i(z)}{\pi'_i}, \quad \widehat{R}(z) = \frac{\widehat{t}_{h(z)_m}}{\widehat{t}_{g(z)_m}},$$
$$\widehat{t}_{h(z)_m} = \sum_{i\in s_m} \frac{h_i(z)}{\pi'_i\pi_{i|s'}}, \quad \widehat{t}_{g(z)_m} = \sum_{i\in s_m} \frac{g_i(z)}{\pi'_i\pi_{i|s'}},$$

which corresponds to the choice $b = \frac{\widehat{t}_{h(z)_m}}{\widehat{t}_{g(z)_m}} = \widehat{R}(z)$.

A second estimator $\widehat{F}_{yu}(z)$ (5) of the distribution function $F_y(z)$ can be obtained from the unmatched sample s_u. By linear combination of $\widehat{F}^r_{ym}(z)$ and $\widehat{F}_{yu}(z)$, we obtain a new composite ratio type estimator

$$\widehat{F}^r_y(z) = \lambda\frac{1}{N}\widehat{t}^r_{h(z)_m} + (1-\lambda)\frac{1}{N}\widehat{t}_{h(z)_u}, \tag{23}$$

where λ is a constant $(0 < \lambda < 1)$.

In the case of simple random sampling on each of the two occasions, the ratio type estimator (23) of the distribution function $F_y(z)$, using the two-occasion scheme is expressed:

$$\widehat{F}^r_y(z) = \lambda\overline{g(z)}_{n'}\widehat{R}(z) + (1-\lambda)\frac{1}{N}\overline{h(z)}_u, \tag{24}$$

where

$$\widehat{R}(z) = \frac{\sum_{i\in s_m} h_i(z)}{\sum_{i\in s_m} g_i(z)},$$

λ is a constant $(0 < \lambda < 1)$, and $\overline{g(z)}_{n'}$, $\overline{h(z)}_u$ have been introduced in (12).

Proposition 4. *In the case of simple random sampling on each of the two occasions, the approximate variance* $\mathrm{AVar}(\widehat{F}^r_y(z))$ *of the ratio type estimator* $\widehat{F}^r_y(z)$ (24) *of the distribution function* $F_y(z)$ *is expressed:*

$$\mathrm{AVar}\big(\widehat{F}^r_y(z)\big) = \lambda^2\frac{1}{N^2}\mathrm{AVar}\big(\widehat{t}^r_{h(z)_m}\big) + (1-\lambda)^2\frac{1}{N^2}\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) + 2\lambda(1-\lambda)\frac{1}{N^2}\mathrm{Cov}\big(\widehat{t}^r_{h(z)_m}, \widehat{t}_{h(z)_u}\big), \tag{25}$$

where

$$\mathrm{AVar}\big(\widehat{t}^r_{h(z)_m}\big) = N^2\left(\left(1-\frac{n'}{N}\right)\frac{s^2_{h(z)}}{n'} + \left(1-\frac{m}{n'}\right)\frac{s^2_{R(z)}}{m}\right), \tag{26}$$

$$s^2_{R(z)} = \frac{1}{N-1}\sum_{i=1}^{N}\big(h_i(z) - R(z)g_i(z)\big)^2, \quad R(z) = \frac{\sum_{i=1}^N h_i(z)}{\sum_{i=1}^N g_i(z)},$$

$$\mathrm{Cov}\big(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u}\big) = -N s^2_{h(z)},$$

$\mathrm{Var}(\widehat{t}_{h(z)_u})$ *and* $s^2_{h(z)}$ *are given in* (13) *and* (6)*.*

Proof. The variance of the composite ratio type estimator (23) equals

$$\begin{aligned}\mathrm{Var}\big(\widehat{F}^r_y(z)\big) = \lambda^2 \frac{1}{N^2}\mathrm{Var}\big(\widehat{t}^{\,r}_{h(z)_m}\big) + (1-\lambda)^2 \frac{1}{N^2}\mathrm{Var}\big(\widehat{t}_{h(z)_u}\big) \\ + 2\lambda(1-\lambda)\frac{1}{N^2}\mathrm{Cov}\big(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u}\big). \end{aligned} \tag{27}$$

The approximation of $\mathrm{Var}(\widehat{t}^{\,r}_{h(z)_m})$ is given:

$$\begin{aligned}\mathrm{AVar}\big(\widehat{t}^{\,r}_{h(z)_m}\big) = \sum_{i,j\in\mathcal{U}} \big(\pi'_{ij} - \pi'_i\pi'_j\big)\frac{h_i(z)}{\pi'_i}\frac{h_j(z)}{\pi'_j} \\ + \mathrm{E}\left(\sum_{i,j\in s'} \big(\pi_{ij|s'} - \pi_{i|s'}\pi_{j|s'}\big)\frac{R_i(z)}{\pi'_i\pi_{i|s'}}\frac{R_j(z)}{\pi'_j\pi_{j|s'}}\right), \end{aligned} \tag{28}$$

with $R_i(z) = h_i(z) - R(z)g_i(z)$. $R(z)$ defined in (24). $\mathrm{Var}(\widehat{t}_{h(z)_u})$ and $\mathrm{Cov}(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u}) = \mathrm{Cov}(\widehat{t}_{h(z)_m}, \widehat{t}_{h(z)_u})$ are given in (16) and (17).

Replacing π values in (27), (28) by the corresponding values, obtained for a simple random sampling design on each of the two occasions, we get an expression of the approximate variance (25) of the distribution function estimator (24). □

Remark 4. We use variance estimator $\widehat{\mathrm{Var}}(\widehat{F}^r_y(z))$ of the composite ratio type distribution function estimator (24) replacing $s^2_{h(z)}$, $s^2_{R(z)}$ in the $\mathrm{AVar}(\widehat{t}^{\,r}_{h(z)_m})$ of (25) by the corresponding estimates

$$\widehat{s}^2_{R(z)_m} = \frac{1}{m-1}\sum_{i\in s_m}\big(h_i(z) - \widehat{R}(z)g_i(z)\big)^2, \quad \widehat{R}(z) = \frac{\sum_{i\in s_m} h_i(z)}{\sum_{i\in s_m} g_i(z)},$$

$\widehat{s}^2_{h(z)_m}$ are given in Remark 2.

The estimators $\widehat{\mathrm{Var}}(\widehat{t}_{h(z)_u})$ and $\widehat{\mathrm{Cov}}(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u}) = \widehat{\mathrm{Cov}}(\widehat{t}^{\,reg}_{h(z)_m}, \widehat{t}_{h(z)_u})$ have been obtained for the variance estimator $\widehat{\mathrm{Var}}(\widehat{F}^{reg}_y(z))$.

Using the same ideas as for obtaining a composite optimal regression type estimator of the distribution function, we obtain a composite optimal ratio type estimator of the distribution function. In the case of simple random sampling for each of the two occasions:

$$\widehat{F}^r_{y\ opt}(z) = \lambda_{opt}\overline{g(z)}_{n'}\widehat{R}(z) + (1-\lambda_{opt})\frac{1}{N}\overline{h(z)}_u, \tag{29}$$

where

$$\lambda_{opt} = \frac{\mathrm{Var}(\widehat{t}_{h(z)_u}) - \mathrm{Cov}(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u})}{\mathrm{Var}(\widehat{t}^{\,r}_{h(z)_m}) + \mathrm{Var}(\widehat{t}_{h(z)_u}) - 2\mathrm{Cov}(\widehat{t}^{\,r}_{h(z)_m}, \widehat{t}_{h(z)_u})}.$$

The approximate minimal variance $\mathrm{AVar}(\widehat{F}^r_{y\ opt}(z))_{min}$ of the ratio type estimator $\widehat{F}^r_{y\ opt}(z)$ (29) of the distribution function $F_y(z)$ is expressed:

$$\mathrm{Var}\big(\widehat{F}^r_{y\ opt}(z)\big)_{min} = \frac{1}{N^2}\left(\frac{\mathrm{Var}_1\mathrm{Var}_2 - \mathrm{Cov}^2}{\mathrm{Var}_1 + \mathrm{Var}_2 - 2\mathrm{Cov}}\right), \tag{30}$$

where $\mathrm{Var}_1 = \mathrm{Var}(\widehat{t}^r_{h(z)_m})$, $\mathrm{Var}_2 = \mathrm{Var}(\widehat{t}_{h(z)_u})$, $\mathrm{Cov} = \mathrm{Cov}(\widehat{t}^r_{h(z)_m}, \widehat{t}_{h(z)_u})$.

The coefficient λ_{opt} depends on unknown variances and covariance, and we have to estimate it by

$$\widehat{\lambda}_{opt} = \frac{\widehat{\mathrm{Var}}(\widehat{t}_{h(z)_u}) - \widehat{\mathrm{Cov}}(\widehat{t}^r_{h(z)_m}, \widehat{t}_{h(z)_u})}{\widehat{\mathrm{Var}}(\widehat{t}^r_{h(z)_m}) + \widehat{\mathrm{Var}}(\widehat{t}_{h(z)_u}) - 2\widehat{\mathrm{Cov}}(\widehat{t}^r_{h(z)_m}, \widehat{t}_{h(z)_u})}. \tag{31}$$

Finally, we use the minimal variance estimator $\widehat{\mathrm{Var}}(\widehat{F}^r_y(z))_{min}$ of the composite optimal ratio type estimator (29) of the distribution function replacing Var_1, Var_2, and Cov in $\mathrm{Var}(\widehat{t}^{reg}_{h(z)_m})_{min}$ of (30) by the corresponding estimators $\widehat{\mathrm{Var}}_1$, $\widehat{\mathrm{Var}}_2$, and $\widehat{\mathrm{Cov}}$.

3 Simulation study

In this section, we present a simulation study for the comparison of the performance of several distribution function estimators using the scheme of two-occasion sampling, with simple random sampling on each of the two occasions.

We study real household data of Statistics Lithuania. The study population consists of $N = 2\,932$ households. The data are available for two occasions. The variables of interest, y and x, are the total household gross income; the values x_i (the first occasion) refer to the population in 2005, the values y_i (the second occasion) refer to the population in 2006. The correlation coefficient between the variables x and y in the household population is $\varrho(x,\ y) = 0.86$. It means a strong linear relationship. To construct the estimator $\widehat{F}_y(z)$, we have chosen the following points z_k:

$$z_1 = K_{0.10}, \quad z_2 = K_{0.25}, \quad z_3 = K_{0.50}, \quad z_4 = K_{0.75}, \quad z_5 = K_{0.90},$$

where K_q is the q-level quantile of the study variable y in the household population.

We have selected $B = 10\,000$ samples of size $n' = 200$ on the first occasion under simple random sampling, with different matching fractions on the second occasion: $\frac{m}{n} = \frac{1}{4}$ ($m = 50, u = 150$), $\frac{m}{n} = \frac{1}{2}$ ($m = 100, u = 100$) and $\frac{m}{n} = \frac{3}{4}$ ($m = 150, u = 50$) under simple random sampling as well. For each sample we compute several estimators of the population distribution function: a simple estimator $\widehat{F}_y(z)$, ratio and regression type estimators $\widehat{F}^r_y(z)$ and $\widehat{F}^{reg}_y(z)$, respectively, with the coefficient $\omega = 0.5$ and $\lambda = 0.5$, as well as optimal ratio and regression type estimators $\widehat{F}^r_{yopt}(z)$ and $\widehat{F}^{reg}_{yopt}(z)$, respectively, in the sense of minimizing variance with the optimal coefficients ω_{opt} and λ_{opt}.

For each estimator, we have calculated estimates of the distribution function of the study variable y at the points $K_{0.10}$, $K_{0.25}$, $K_{0.50}$, $K_{0.75}$, and $K_{0.90}$. Thus, for each

estimator we have 10 000 estimates of $\widehat{F}_y(z)$ for z=$K_{0.10}$, $K_{0.25}$, $K_{0.50}$, $K_{0.75}$, and $K_{0.90}$. For these estimates we have calculated a relative bias, relative root mean square errors, and a relative efficiency of the estimators. For each estimator $\widehat{\theta}_y(z)$ we define the relative bias as

$$\mathrm{RB}\big(\widehat{\theta}_y(z)\big) = \frac{1}{F_y(z)}\left(\frac{1}{B}\sum_{i=1}^{B}\big(\widehat{\theta}_y^{(i)}(z) - F_y(z)\big)\right),$$

the relative root mean square error as

$$\mathrm{RMSE}\big(\widehat{\theta}_y(z)\big) = \frac{1}{F_y(z)}\sqrt{\frac{1}{B}\sum_{i=1}^{B}\big(\widehat{\theta}_y^{(i)}(z) - F_y(z)\big)^2},$$

where $\widehat{\theta}_y^{(i)}(z)$ is the i-th estimate at the point z, calculated for the estimator $\widehat{\theta}_y(z)$, the relative efficiency with

$$\mathrm{RE}\big(\widehat{\theta}_y(z)\big) = \frac{\mathrm{RMSE}(\widehat{F}_y(z))}{\mathrm{RMSE}(\widehat{\theta}_y(z))},$$

where $\mathrm{RMSE}(\widehat{F}_y(z))$ is the relative root mean square error defined for the simple estimator $\widehat{F}_y(z)$. For each estimator also we define efficiency $\mathrm{E}(\widehat{\theta}_y(z))$, the ratio of the $\mathrm{RMSE}(\widehat{F}_y(z))$ and $\mathrm{RMSE}(\widehat{\theta}_y(z))$ with the corresponding formulae based variances.

Table 1 illustrates the relative bias of proposed estimators of the population distribution function. For non-optimal estimators of a distribution function, the relative bias is decreasing when the q level of the population quantile is increasing. The simple and regression type estimators, $\widehat{F}_y(z)$ and $\widehat{F}_y^{reg}(z)$, respectively, in all cases behave in a similar way, but in most cases $\widehat{F}_y^{reg}(z)$ has the lowest relative bias. The regression type estimator $\widehat{F}_y^{reg}(z)$ is less biased than the ratio type estimator $\widehat{F}_y^{r}(z)$, especially for a low q level of the population quantile and for a low matching fraction m/n. The optimal ratio and regression type estimators, $\widehat{F}_{yopt}^{r}(z)$ and $\widehat{F}_{yopt}^{reg}(z)$, respectively, have the highest bias in most cases.

Table 2 shows a relative root mean square error of estimators for the real household population and several matching fractions on the second occasion. As to the efficiency, measured by the relative root mean square error, the regression type estimator $\widehat{F}_y^{reg}(z)$ is more efficient than the ratio type estimator $\widehat{F}_y^{r}(z)$, especially for a low q level of the population quantile and a low matching fraction. This is probably due to the specificity of variables $g(z)$ and $h(z)$. In most cases, a simple estimator $\widehat{F}_y(z)$ of the population distribution function has a high relative root mean square error especially for a low matching fraction and a high q level of the population quantile. The estimators $\widehat{F}_{yopt}^{r}(z)$ and $\widehat{F}_{yopt}^{reg}(z)$ are usually more efficient in most cases, than $\widehat{F}_y^{r}(z)$ and $\widehat{F}^{reg}(z)$, respectively.

The relative efficiency for the proposed estimators $\widehat{F}_y^{r}(z)$, $\widehat{F}_y^{reg}(z)$, $\widehat{F}_{yopt}^{r}(z)$, $\widehat{F}_{yopt}^{reg}(z)$ and for the simple estimator $\widehat{F}_y(z)$ of the population distribution function, using a two-occasion scheme, is presented in Table 3. The estimators $\widehat{F}_{yopt}^{r}(z)$ and $\widehat{F}_{yopt}^{reg}(z)$ are

Table 1. Relative bias (RB) of estimators

	Estimator	$K_{0.10}$	$K_{0.25}$	$K_{0.50}$	$K_{0.75}$	$K_{0.90}$
$\frac{m}{n}=\frac{1}{4}$						
	$\widehat{F}_y$	0.01148	0.00009	−0.00052	−0.00038	−0.00041
	$\widehat{F}_y^r$	0.02512	0.00394	0.00086	0.00019	−0.00010
	$\widehat{F}_y^{reg}$	0.01131	0.00077	−0.00009	−0.00013	−0.00046
	$\widehat{F}_{yopt}^r$	−0.01278	−0.00287	0.00148	0.00348	0.00376
	$\widehat{F}_{yopt}^{reg}$	−0.02036	−0.00478	0.00117	0.00361	0.00403
$\frac{m}{n}=\frac{1}{2}$						
	$\widehat{F}_y$	0.01429	−0.00055	−0.00025	0.00015	0.00015
	$\widehat{F}_y^r$	0.01614	0.00171	0.00063	0.00031	0.00014
	$\widehat{F}_y^{reg}$	0.01504	0.00046	0.00027	0.00022	0.00012
	$\widehat{F}_{yopt}^r$	−0.01740	−0.00546	0.00082	0.00300	0.00379
	$\widehat{F}_{yopt}^{reg}$	−0.01804	−0.00689	0.00049	0.00298	0.00393
$\frac{m}{n}=\frac{3}{4}$						
	$\widehat{F}_y$	0.01636	0.00112	0.00135	0.00025	0.00045
	$\widehat{F}_y^r$	0.01682	0.00153	0.00146	0.00023	0.00040
	$\widehat{F}_y^{reg}$	0.01560	0.00129	0.00141	0.00024	0.00040
	$\widehat{F}_{yopt}^r$	−0.02234	−0.00827	0.00105	0.00313	0.00500
	$\widehat{F}_{yopt}^{reg}$	−0.02415	−0.00865	0.00098	0.00315	0.00502

Table 2. Relative root mean square error (RMSE) of estimators

	Estimator	$K_{0.10}$	$K_{0.25}$	$K_{0.50}$	$K_{0.75}$	$K_{0.90}$
$\frac{m}{n}=\frac{1}{4}$						
	$\widehat{F}_y$	0.2066	0.1206	0.0688	0.0397	0.0233
	$\widehat{F}_y^r$	0.2340	0.1192	0.0667	0.0393	0.0233
	$\widehat{F}_y^{reg}$	0.2160	0.1171	0.0659	0.0392	0.0299
	$\widehat{F}_{yopt}^r$	0.2207	0.1136	0.0639	0.0378	0.0299
	$\widehat{F}_{yopt}^{reg}$	0.2173	0.1131	0.0637	0.0381	0.0233
$\frac{m}{n}=\frac{1}{2}$						
	$\widehat{F}_y$	0.2083	0.1195	0.0693	0.0401	0.0299
	$\widehat{F}_y^r$	0.1996	0.1115	0.0642	0.0376	0.0214
	$\widehat{F}_y^{reg}$	0.1969	0.1110	0.0640	0.0375	0.0214
	$\widehat{F}_{yopt}^r$	0.2065	0.1119	0.0638	0.0378	0.0223
	$\widehat{F}_{yopt}^{reg}$	0.2040	0.1113	0.0636	0.0378	0.0224
$\frac{m}{n}=\frac{3}{4}$						
	$\widehat{F}_y$	0.2069	0.1194	0.0687	0.0394	0.0228
	$\widehat{F}_y^r$	0.2377	0.1371	0.0787	0.0455	0.0264
	$\widehat{F}_y^{reg}$	0.2368	0.1370	0.0786	0.0454	0.0263
	$\widehat{F}_{yopt}^r$	0.2192	0.1159	0.0655	0.0385	0.0244
	$\widehat{F}_{yopt}^{reg}$	0.2175	0.1155	0.0654	0.0385	0.0244

Table 3. Relative efficiency (RE) of estimators

	Estimator	$K_{0.10}$	$K_{0.25}$	$K_{0.50}$	$K_{0.75}$	$K_{0.90}$
$\frac{m}{n}=\frac{1}{4}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	0.883	1.012	1.032	1.012	1.003
	$\widehat{F}_y^{reg}$	0.956	1.030	1.043	1.014	1.017
	$\widehat{F}_{yopt}^r$	0.936	1.062	1.076	1.050	1.020
	$\widehat{F}_{yopt}^{reg}$	0.951	1.067	1.079	1.044	0.999
$\frac{m}{n}=\frac{1}{2}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	1.044	1.072	1.080	1.067	1.070
	$\widehat{F}_y^{reg}$	1.058	1.077	1.083	1.069	1.070
	$\widehat{F}_{yopt}^r$	1.009	1.068	1.087	1.060	1.027
	$\widehat{F}_{yopt}^{reg}$	1.021	1.074	1.090	1.062	1.021
$\frac{m}{n}=\frac{3}{4}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	0.871	0.871	0.873	0.867	0.866
	$\widehat{F}_y^{reg}$	0.874	0.872	0.874	0.868	0.867
	$\widehat{F}_{yopt}^r$	0.944	1.031	1.048	1.023	0.936
	$\widehat{F}_{yopt}^{reg}$	0.952	1.034	1.051	1.025	0.937

Table 4. Efficiency (E) of estimators

	Estimator	$K_{0.10}$	$K_{0.25}$	$K_{0.50}$	$K_{0.75}$	$K_{0.90}$
$\frac{m}{n}=\frac{1}{4}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	0.781	1.004	1.075	1.017	1.002
	$\widehat{F}_y^{reg}$	0.995	1.074	1.113	1.043	1.039
	$\widehat{F}_{yopt}^r$	1.133	1.163	1.195	1.171	1.219
	$\widehat{F}_{yopt}^{reg}$	1.187	1.196	1.213	1.183	1.237
$\frac{m}{n}=\frac{1}{2}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	1.092	1.148	1.170	1.145	1.139
	$\widehat{F}_y^{reg}$	1.126	1.163	1.179	1.151	1.150
	$\widehat{F}_{yopt}^r$	1.141	1.180	1.204	1.177	1.206
	$\widehat{F}_{yopt}^{reg}$	1.183	1.200	1.216	1.185	1.221
$\frac{m}{n}=\frac{3}{4}$						
	$\widehat{F}_y$	1.000	1.000	1.000	1.000	1.000
	$\widehat{F}_y^r$	0.755	0.761	0.764	0.761	0.760
	$\widehat{F}_y^{reg}$	0.758	0.762	0.765	0.762	0.761
	$\widehat{F}_{yopt}^r$	1.119	1.121	1.129	1.117	1.152
	$\widehat{F}_{yopt}^{reg}$	1.135	1.128	1.134	1.121	1.159

as usual more efficient than $\widehat{F}_y^r(z)$ and $\widehat{F}_y^{reg}(z)$, respectively, in most cases. Optimal estimators mostly have a higher bias shown before. That is a reason why sometimes the efficiency is decreasing, compared with other estimators. The relative efficiency of a simple estimator is higher for the lowest q level of the population quantile with a lowest and highest matching fraction. The relative efficiency of the optimal distribution function estimators at the median are highest with any sampling fractions.

The efficiency, ratio of the RMSE with the corresponding formulae based variances $\widehat{\mathrm{Var}}(\widehat{F}_y^r(z))$, $\widehat{\mathrm{Var}}(\widehat{F}_y^{reg}(z))$, $\widehat{\mathrm{Var}}(\widehat{F}_{yopt}^r(z))$, $\widehat{\mathrm{Var}}(\widehat{F}_{yopt}^{reg}(z))$ and $\widehat{\mathrm{Var}}(\widehat{F}_y(z))$ of the population distribution function, using a two-occasion scheme, is presented in Table 4. Efficiency of proposed optimal estimators using ratio of RMSE with the corresponding formulae based variance is grows up comparable with relative efficiency. Average estimates of the variances of the proposed optimal ratio and regression estimators are smaller than the empirical variances. The Taylor series expansion of the ratio and regression estimators are used for the expressions of approximate variances. If higher order terms of Taylor expansion would be taken into expression of the approximate variances of these estimators, one can expect to improve the accuracy of the approximation of the variances.

4 Conclusions

We have proposed composite regression and ratio type estimators for a distribution function, as well as optimal estimators, in the sense of minimizing the variance for a two-occasion sampling scheme with a simple random sampling design on each occasion. Simulation has been studied on the real population of Lithuanian households of Statistics Lithuania. The simulation results show that the proposed composite estimators using auxiliary information can be used for improving the accuracy of distribution function estimates. The efficiency of the estimators proposed depends on the matching fraction and on the level of quantiles for two-occasion sampling.

References

1. V. Chadyšas, Estimation of confidence intervals for quantiles in a finite population, *Math. Model. Anal.*, **13**(2), pp. 195–202, 2008.

2. C. E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.

3. R. L. Chambers, R. Dunstan, Estimating distribution functions from survey data, *Biometrika*, **73**, pp. 597–604, 1986.

4. A. H. Dorfman, P. Hall, A comparison of design-based and model-based estimators of the finite population distribution function, *Aust. J. Stat.*, **35**, pp. 29–41, 1993.

5. J. N. K. Rao, J. G. Kovar, H. J. Mantel, On estimating distribution functions and quantiles from survey data using auxiliary information, *Biometrika*, **77**, pp. 365–375, 1990.

6. M. Rueda, S. Martinez, H. Martinez, A. Arcos, Estimation of the distribution function with calibration methods, *J. Stat. Plan. Infer.*, **137**, pp. 435–448, 2007.

7. P. L. D. N. Silva, C J. Skinner, Estimating distribution functions with auxiliary information using poststratification, *Journal of Official Statistics*, **11**, pp. 277–294, 1995.

8. W. G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, New York, 1977.

9. G. Kulldorff, Some problems of optimum allocation for sampling on two occasions, *Rev. Inst. Int. Stat.*, **31**(1), pp. 24–57, 1963.

10. H. D. Patterson, Sampling on successive occasions with partial replacement of units, *J. Roy. Stat. Soc. B Met.*, **12**, pp. 241–255, 1950.

11. M. Rueda, J. F. Munoz, S. Gonzales, A. Arcos, Estimating quantiles under sampling on two occasions with arbitrary sample designs, *Comput. Stat. Data An.*, **51**, pp. 6956–6613, 2007.