Nonlinear Analysis: Modelling and Control, Vol. 21, No. 2, 241–260
http://dx.doi.org/10.15388/NA.2016.2.7 ISSN 1392-5113

Estimation of the current population total for a four-phase sampling design

Viktoras Chadyšas, Danutė Krapavickaitė

Vilnius Gediminas Technical University,
Saulėtekio ave. 11, LT-10283 Vilnius, Lithuania
viktoras.chadysas@vgtu.lt; danute.krapavickaite@vgtu.lt

Received: October 1, 2014 / **Revised:** September 1, 2015 / **Published online:** December 30, 2015

Abstract. The combined ratio-type estimators of the finite population total and their variances in the case of sample rotation for two-phase and four-phase sampling schemes are constructed in the paper. Combined estimators of the finite population total without and with the use of auxiliary information known from the previous survey are built. Two types of sampling design are used for sample selection in each of the phases: simple random sampling without replacement and successive sampling without replacement with probabilities proportional to size. A simulation study, based on the real data, is performed, and the accuracy of the estimators proposed is compared.

Keywords: sample rotation, successive sampling, ratio estimator, composite estimator.

1 Introduction

A sample survey when information is collected regularly on the same population in subsequent time periods with the partial replacement of the sample is studied. Such repetitive surveys (consecutive measurements of the same population) are used in social studies, official statistics, forestry, medicine etc. The Labour Force Survey (LFS) provides estimates of the number of employed and unemployed individuals for each quarter of the year. Repeated sampling from a finite population (or sample rotation) is a sampling procedure, which is usually used for this survey.

Let us denote a finite household population $\mathcal{U} = \{1, \ldots, i, \ldots, N\}$ of size N. For each household, the number of its members is denoted by m_i, $i = 1, 2, \ldots, N$. The sum of household members is obtained by $M = \sum_{i=1}^{N} m_i$. Let us suppose that the survey variable y means the number of employed (or unemployed) individuals in each household. The values y_i of the variable belong to the set of integers $\{0, 1, \ldots, m_i\}$. The parameter of interest is the total of the number of employed (or unemployed) individuals

$$t_y = \sum_{i=1}^{N} y_i. \tag{1}$$

The previous survey data can be used as auxiliary information for the estimation of the population total in order to reduce the variance of the estimator. The efficiency of ratio estimators in the case of any sampling design is discussed in Särndal et al. [21]. Combined estimators of the finite population total without and with the use of auxiliary information known from the previous survey are constructed.

Sample rotation and two-phase sampling (double sampling) are similar procedures. Sample rotation means that a sample of the current occasion consists of a union of subsamples: one of them is matched with the elements of previous occasions, and the other one is a new one and unmatched with the previously studied elements. A sample under a two-phase sampling design is matched with the first-phase sample.

In this paper, the construction of the combined estimator of the finite population total and its variance in the case of sample rotation is analyzed for two-phase and four-phase sampling schemes.

If the auxiliary variable is well-correlated with the study variable, then it is possible to obtain more accurate estimates of the parameter. A two-phase sampling design and estimators of the total with the use of auxiliary information are given in Särndal et al. [21]. They are applied here to the Lithuanian LFS data in the case of a simple random sample of individuals, and the current paper is a further development of [7].

The estimators for a total in the case of the three-phase sampling design are presented in Fuller [9] and Singh [22]. Jeyaratnam et al. [13] studied multiphase sampling for the stratification and efficient allocation of the sample size. Many various problems are being solved for two occasions sample data. One of them is the optimal choice of the second-occasion sampling design in order to minimize the variance of the estimator (Arnab, [3]). Hamad et al. [10] uses two auxiliary variables for the estimator of the total in a two-phase sampling design. Subsampling of nonrespondents and corresponding estimators is also a case of estimation under a two-phase sampling design (Okafor and Lee, [16]), under a three-phase sampling design (Hidiriglou and Estevao [11]). Artes and Garcia [4] studied the estimators of the ratio under sampling on two occasions with partial replacement of the elements. Close attention is paid to variance estimation for the estimator of change in the finite population parameter in repeated surveys. There are many studies on this topic, for example, Berger [5], Andersson [2], and Qualité [17]. Fattorini et al. [8] studied a special three-phase sampling strategy for the estimation of forest biomass.

Combined estimators of the population total are obtained taking a linear combination of ratio estimators using the π^* estimators idea (Särndal et al., [21]) for a multi-phase sampling design and the Horvitz–Thompson estimator (Horvitz and Thompson, [12]) or the ratio estimator. Two types of sampling design are used here for sample selection in each of the phases: simple random sampling without replacement and a successive sampling (unequal probability sampling without replacement) procedure proposed by Rosén [18]. The second-order inclusion probabilities for a successive sampling design are approximated by corresponding probabilities for conditional Poisson sampling. The results of Aires [1] and Bondesson et al. [6] are used for this. Then the inclusion probabilities obtained are used to calculate the estimates of the proposed estimators of the totals and their variance estimates. A simulation study, based on the real population data, is performed, and the estimators proposed are compared.

2 Sample rotation and sample selection

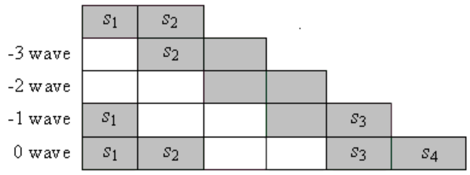
The LFS at Statistics Lithuania is conducted continuously with a quarterly selected sample. All members of a household are included in the sample for two subsequent quarters, excluded from the sample for the next two quarters, and included once more in the sample for two other quarters. It means that one-fourth of the sample of the previous quarter is replaced by the new one each quarter of the year as shown in Fig. 1.

The sample selection procedure is performed as shown in Fig. 2.

It is seen in the sample selection scheme, presented in Fig. 2, that the whole sample s consists of a union of four subsamples: s_1, s_2, s_3 and s_4. The subsamples selected at each of the phases are expressed:

$$
\begin{aligned}
&s_1 : \mathcal{U} \to s_1;\\
&s_2 : \mathcal{U} \to s_2' = \mathcal{U} \setminus s_1 \to s_2, \text{ two phases;}\\
&s_3 : \mathcal{U} \to s_2' = \mathcal{U} \setminus s_1 \to s_3' = \mathcal{U} \setminus (s_1 \cup s_2) \to s_3, \text{ three phases;}\\
&s_4 : \mathcal{U} \to s_2' = \mathcal{U} \setminus s_1 \to s_3' = \mathcal{U} \setminus (s_1 \cup s_2)\\
&\qquad \to s_4' = \mathcal{U} \setminus (s_1 \cup s_2 \cup s_3) \to s_4, \text{ four phases.}
\end{aligned}
$$

The estimators under the sampling scheme described will be discussed further.



	s_1	s_2				
-3 wave		s_2	▨			
-2 wave			▨	▨		
-1 wave	s_1			▨	s_3	
0 wave	s_1	s_2			s_3	s_4

Figure 1. Sample rotation scheme of the Labour Force Survey.

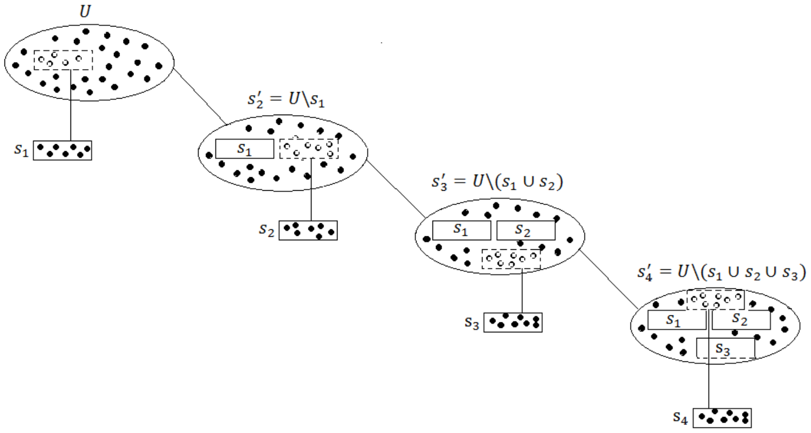

U
s_1
$s_2' = U\backslash s_1$
s_1
s_2
$s_3' = U\backslash(s_1 \cup s_2)$
s_1
s_2
s_3
$s_4' = U\backslash(s_1 \cup s_2 \cup s_3)$
s_1
s_2
s_3
s_4

Figure 2. Labour Force Survey four-phase sampling scheme.

3 Simple estimators of the population total

We are interested in the estimation of the population total $t_y = \sum_{i \in U} y_i$ for a study variable y. Firstly, we construct four separate design-based estimators of the total using data of the samples s_1, s_2, s_3 and s_4, respectively. Secondly, we propose a combined estimator of the total using sample rotation schemes in Section 4.

Step 1. The sample s_1 is selected from the finite population: $\mathcal{U} \to s_1$. The corresponding first- and second-order inclusion probabilities for elements of the sample s_1 are denoted respectively:

$$\pi_{1i} = \mathbf{P}(i \in s_1); \quad \pi_{1ij} = \mathbf{P}(i \in s_1, j \in s_1), \quad \pi_{1ii} = \pi_{1i}.$$

An unbiased design-based Narain [15] and Horvitz–Thompson [12] estimator of the population total is used:

$$\hat{t}_{1y}^{\mathrm{HT}} = \sum_{i \in s_1} \frac{y_i}{\pi_{1i}}. \tag{2}$$

The variance of the estimator $\hat{t}_{1y}^{\mathrm{HT}}$ and its unbiased estimator is

$$\operatorname{Var}\bigl(\hat{t}_{1y}^{\mathrm{HT}}\bigr) = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}}, \tag{3}$$

$$\widehat{\operatorname{Var}}\bigl(\hat{t}_{1y}^{\mathrm{HT}}\bigr) = \sum_{i \in s_1} \sum_{j \in s_1} \left(1 - \frac{\pi_{1i}\pi_{1j}}{\pi_{1ij}}\right) \frac{y_i}{\pi_{1i}} \frac{y_j}{\pi_{1j}}. \tag{4}$$

The values of the study variable y in the previous survey can be used as auxiliary information. Let us denote the study variable of the previous survey (-1 wave) by x with the values x_i and the same variable on the current 0 wave by y with the values y_i, $i \in s_1$. We can form the ratio estimator $\hat{t}_{1y}^{\mathrm{rat}}$ of the population total t_y by

$$\hat{t}_{1y}^{\mathrm{rat}} = t_{1x}^{(-1)} \frac{\hat{t}_{1y}}{\hat{t}_{1x}} = t_{1x}^{(-1)} \hat{r}, \quad \hat{r} = \frac{\hat{t}_{1y}}{\hat{t}_{1x}}, \tag{5}$$

We use here $\hat{t}_{1y} = \hat{t}_{1y}^{\mathrm{HT}}$, $\hat{t}_{1x} = \hat{t}_{1x}^{\mathrm{HT}}$, $t_{1x}^{(-1)} = \sum_{i \in \mathcal{U}} x_i$. Some other approximately unbiased estimators $\hat{t}_{1y}$, $\hat{t}_{1x}$ will be also used in this situation further. The estimator $\hat{t}_{1y}^{\mathrm{rat}}$ is nonlinear. Its approximate variance based on a Taylor linearization of the estimator is expressed as:

$$\operatorname{AVar}\bigl(\hat{t}_{1y}^{\mathrm{rat}}\bigr) = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} (\pi_{1ij} - \pi_{1i}\pi_{1j}) \frac{y_i - r x_i}{\pi_{1i}} \frac{y_j - r x_j}{\pi_{1j}} \tag{6}$$

with $r = \sum_{i \in \mathcal{U}} y_i / \sum_{i \in \mathcal{U}} x_i$. The variance $\operatorname{AVar}(\hat{t}_{1y}^{\mathrm{rat}})$ is estimated by

$$\widehat{\operatorname{Var}}\bigl(\hat{t}_{1y}^{\mathrm{rat}}\bigr) = \sum_{i \in s_1} \sum_{j \in s_1} \left(1 - \frac{\pi_{1i}\pi_{1j}}{\pi_{1ij}}\right) \frac{y_i - \hat{r} x_i}{\pi_{1i}} \frac{y_j - \hat{r} x_j}{\pi_{1j}}, \tag{7}$$

using $\hat{r}$ given in (5).

Step 2. The sample s_2 is obtained in two-phase sampling: $\mathcal{U} \to s_2' = \mathcal{U} \setminus s_1 \to s_2$. The corresponding first- and second-order unconditional and conditional element inclusion probabilities for samples s_2' (first phase) and s_2 (second phase) are, respectively,

$$\begin{aligned} \pi_{2ij}' &= \mathbf{P}(i \in s_2', j \in s_2') \\ &= 1 - \mathbf{P}(i \in s_1, j \in s_1) - \mathbf{P}(i \in s_1, j \notin s_1) - \mathbf{P}(i \notin s_1, j \in s_1), \\ \pi_{2i}' &= \mathbf{P}(i \in s_2') = \mathbf{P}(i \notin s_1) = 1 - \mathbf{P}(i \in s_1), \\ \pi_{2i|s_2'} &= \mathbf{P}(i \in s_2 \mid s_2'), \\ \pi_{2ij|s_2'} &= \mathbf{P}(i \in s_2, j \in s_2 \mid s_2'). \end{aligned} \tag{8}$$

Under a two-phase sampling design, using the π^* estimator defined in [21, Sect. 9.2], the population total t_y is unbiasedly estimated by

$$\hat{t}_{2y}^{(2)} = \sum_{i \in s_2} \frac{y_i}{\pi_{2i}' \pi_{2i|s_2'}}. \tag{9}$$

In the case of two-phase sampling, the variance of the estimator $\hat{t}_{2y}^{(2)}$ may be expressed by conditional and unconditional variances and expectations:

$$\begin{aligned} \operatorname{Var}\bigl(\hat{t}_{2y}^{(2)}\bigr) &= \operatorname{Var}\bigl\{\mathbf{E}\bigl(\hat{t}_{2y}^{(2)} \mid s_2'\bigr)\bigr\} + \mathbf{E}\bigl\{\operatorname{Var}\bigl(\hat{t}_{2y}^{(2)} \mid s_2'\bigr)\bigr\} \\ &= \operatorname{Var}\bigl(\hat{t}_{2y}^{(1)}\bigr) + \mathbf{E}\bigl\{\operatorname{Var}\bigl(\hat{t}_{2y}^{(2)} \mid s_2'\bigr)\bigr\}, \end{aligned} \tag{10}$$

$$\hat{t}_{2y}^{(1)} = \sum_{i \in s_2'} \frac{y_k}{\pi_{2i}'} = \sum_{i \in \mathcal{U} \setminus s_1} \frac{y_k}{\pi_{2i}'}$$

or

$$\begin{aligned} \operatorname{Var}\bigl(\hat{t}_{2y}^{(2)}\bigr) &= \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} \bigl(\pi_{2ij}' - \pi_{2i}'\pi_{2j}'\bigr) \frac{y_i}{\pi_{2i}'} \frac{y_j}{\pi_{2j}'} \\ &\quad + \mathbf{E} \sum_{i,j \in s_2'} \bigl(\pi_{2ij|s_2'} - \pi_{2i|s_2'}\pi_{2j|s_2'}\bigr) \frac{y_i}{\pi_{2i}'\pi_{2i|s_2'}} \frac{y_j}{\pi_{2j}'\pi_{2j|s_2'}}. \end{aligned} \tag{11}$$

The variance $\operatorname{Var}(\hat{t}_{2y}^{(2)})$ is estimated unbiasedly by

$$\begin{aligned} \widehat{\operatorname{Var}}\bigl(\hat{t}_{2y}^{(2)}\bigr) &= \sum_{i,j \in s_2} \frac{\pi_{2ij}' - \pi_{2i}'\pi_{2j}'}{\pi_{2ij}'\pi_{2ij|s_2'}} \frac{y_i}{\pi_{2i}'} \frac{y_j}{\pi_{2j}'} \\ &\quad + \sum_{i,j \in s_2} \frac{\pi_{2ij|s_2'} - \pi_{2i|s_2'}\pi_{2j|s_2'}}{\pi_{2ij|s_2'}} \frac{y_i}{\pi_{2i}'\pi_{2i|s_2'}} \frac{y_j}{\pi_{2j}'\pi_{2j|s_2'}}. \end{aligned} \tag{12}$$

The values of the study variable y in the previous survey (-3 wave) can be used as auxiliary information. Let us denote the study variable of the previous survey by x with the values x_i and the same variable on the current wave by y with the values y_i, $i \in s_2$. We can form a ratio estimator $\hat{t}_{2y}^{\mathrm{rat}}$ of the population total t_y by

$$\hat{t}_{2y}^{\mathrm{rat}} = t_{2x}^{(-3)} \hat{r}_2, \quad \hat{r}_2 = \frac{\hat{t}_{2y}^{(2)}}{\hat{t}_{2x}^{(2)}}, \quad \hat{t}_{2x}^{(2)} = \sum_{i \in s_2} \frac{x_i}{\pi'_{2i} \pi_{2i|s'_2}}. \tag{13}$$

Here $t_{2x}^{(-3)} = \sum_{i \in \mathcal{U}} x_i$ is the total of the variable x, which was a study variable y in the previous survey (-3 wave). The estimator $\hat{t}_{2y}^{(2)}$ is given by (9).

In the case of two-phase sampling, the variance $\mathrm{Var}(\hat{t}_{2y}^{\mathrm{rat}})$ of the estimator $\hat{t}_{2y}^{\mathrm{rat}}$ also may be expressed by conditional and unconditional variances and expectations replacing the estimator $\hat{t}_{2y}^{(2)}$ by the estimator $\hat{t}_{2y}^{\mathrm{rat}}$ in (10). Because $\hat{t}_{2y}^{\mathrm{rat}}$ is a nonlinear estimator, the approximate variance $\mathrm{AVar}(\hat{t}_{2y}^{\mathrm{rat}})$ of $\mathrm{Var}(\hat{t}_{2y}^{\mathrm{rat}})$ is derived using a linear term of its Taylor expansion, and the approximate variance of the ratio estimator $\hat{t}_{2y}^{\mathrm{rat}}$ is

$$\begin{aligned} \mathrm{AVar}\big(\hat{t}_{2y}^{\mathrm{rat}}\big) = {} & \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{U}} (\pi'_{2ij} - \pi'_{2i}\pi'_{2j}) \frac{y_i}{\pi'_{2i}} \frac{y_j}{\pi'_{2j}} \\ & + \mathbf{E} \sum_{i,j \in s'_2} \big(\pi_{2ij|s'_2} - \pi_{2i|s'_2}\pi_{2j|s'_2}\big) \frac{y_i - rx_i}{\pi'_{2i}\pi_{2i|s'_2}} \frac{y_j - rx_j}{\pi'_{2j}\pi_{2j|s'_2}} \end{aligned} \tag{14}$$

with the ratio r given in (6). As the estimator of the variance will be used

$$\begin{aligned} \widehat{\mathrm{Var}}\big(\hat{t}_{2y}^{\mathrm{rat}}\big) = {} & \sum_{i,j \in s_2} \frac{\pi'_{2ij} - \pi'_{2i}\pi'_{2j}}{\pi'_{2ij}\pi_{2ij|s'_2}} \frac{y_i}{\pi'_{2i}} \frac{y_j}{\pi'_{2j}} \\ & + \sum_{i,j \in s_2} \frac{\pi_{2ij|s'_2} - \pi_{2i|s'_2}\pi_{2j|s'_2}}{\pi_{2ij|s'_2}} \frac{y_i - \hat{r}x_i}{\pi'_{2i}\pi_{2i|s'_2}} \frac{y_j - \hat{r}x_j}{\pi'_{2j}\pi_{2j|s'_2}} \end{aligned} \tag{15}$$

with the estimator of the ratio $\hat{r}_2$ given in (13).

Step 3. The sample s_3 is obtained in three-phase sampling: $\mathcal{U} \to s'_2 = \mathcal{U} \setminus s_1 \to s'_3 = \mathcal{U} \setminus (s_1 \cup s_2) \to s_3$. The corresponding first- and second-order inclusion probabilities for the sample s'_2 (first phase) were denoted in Step 2, and for samples s'_3 (second phase) and s_3 (third phase), the inclusion probabilities are, respectively,

$$\pi'_{3i|s'_2} = \mathbf{P}(i \in s'_3 | s'_2) \quad \text{and} \quad \pi_{3i|s'_3} = \mathbf{P}(i \in s_3 \mid s'_3).$$

Under a three-phase sampling design, using the π^* estimator, the population total t_y is unbiasedly estimated by

$$\hat{t}_{3y}^{(3)} = \sum_{i \in s_3} \frac{y_i}{\pi'_{2i}\pi'_{3i|s'_2}\pi_{3i|s'_3}}. \tag{16}$$

In this step, the values of the study variable y in the previous survey (-1 wave) can be used as auxiliary information. Let us denote the study variable in the previous survey by x with the values x_i, and the same variable in the current wave by y with the values y_i, $i \in s_3$. We can form a ratio estimator $\hat{t}_{3y}^{\mathrm{rat}}$ of the population total t_y:

$$\hat{t}_{3y}^{\mathrm{rat}} = \hat{t}_{3x}^{(-1)} \hat{r}_3, \quad \hat{r}_3 = \frac{\hat{t}_{3y}^{(3)}}{\hat{t}_{3x}^{(3)}}, \quad \hat{t}_{3x}^{(3)} = \sum_{i \in s_3} \frac{x_i}{\pi'_{2i} \pi'_{3i|s'_2} \pi_{3i|s'_3}}. \tag{17}$$

Here $t_{3x}^{(-1)} = \sum_{k \in \mathcal{U}} x_k$ is the total of the variable y in the -1 wave. The estimator $\hat{t}_{3y}^{(3)}$ is given in (16).

Step 4. The sample s_4 is obtained in four-phase sampling: $\mathcal{U} \to s'_2 = \mathcal{U} \setminus s_1 \to s'_3 = \mathcal{U} \setminus (s_1 \cup s_2) \to s'_4 = \mathcal{U} \setminus (s_1 \cup s_2 \cup s_3) \to s_4$. The corresponding first- and second-order inclusion probabilities for sample s'_2 (first phase) and s'_3 (second phase) were described in Step 2 and Step 3 previously, and for samples s'_4 (third phase) and s_4 (fourth phase), they are, respectively,

$$\begin{aligned}
\pi'_{4i|s'_3} &= \mathbf{P}(i \in s'_4 \mid s'_3) \\
&= 1 - \mathbf{P}(i \in s_1) - \mathbf{P}(i \in s_2 \mid s'_2)\mathbf{P}(i \in s'_2) - \mathbf{P}(i \in s_3 \mid s'_3)\mathbf{P}(i \in s'_3), \\
\pi_{4i|s'_4} &= \mathbf{P}(i \in s_4 \mid s'_4).
\end{aligned}$$

Under four-phase sampling, using the π^* estimator, the population total t_y is unbiasedly estimated by

$$\hat{t}_{4y}^{(4)} = \sum_{i \in s_4} \frac{y_i}{\pi'_{2i} \pi'_{3i|s'_2} \pi'_{4i|s'_3} \pi_{4i|s'_4}}. \tag{18}$$

More complex estimators are presented further.

4 Combined estimators of the population total

The construction of the combined estimators and their variances of the finite population total (1) in the case of sample rotation for two-phase and four-phase sampling schemes is presented in this section.

1. By a linear combination of $\hat{t}_{1y}$ and $\hat{t}_{2y}^{(2)}$, we obtain the estimator without the use of auxiliary information of the total

$$\hat{t}_2 = \frac{1}{2}\big(\hat{t}_{1y} + \hat{t}_{2y}^{(2)}\big). \tag{19}$$

The expression for the variance of estimator (19) of the total:

$$\begin{aligned}
\mathrm{Var}\big(\hat{t}_2\big) &= \frac{1}{4}\big(\mathrm{Var}(\hat{t}_{1y}) + \mathrm{Var}\big(\hat{t}_{2y}^{(2)}\big) + 2\,\mathrm{Cov}\big(\hat{t}_{1y}, \hat{t}_{2y}^{(2)}\big)\big), \qquad (20) \\
\mathrm{Cov}\big(\hat{t}_{1y}, \hat{t}_{2y}^{(2)}\big) &= \sum_{k \in \mathcal{U}} \sum_{\substack{l \in \mathcal{U} \\ l \neq k}} \frac{y_k}{\pi_{1k}} \frac{y_l}{\pi'_{2l}} \big(\pi_{1k} - \pi_{1kl}^{(1)}\big) - t_y^2,
\end{aligned}$$

$\pi_{1kl}^{(1)} = P(k \in s_1, l \in s_1)$. The variance $\mathrm{Var}(\hat{t}_2)$ is estimated unbiasedly by

$$\widehat{\mathrm{Var}}(\hat{t}_2) = \frac{1}{4}\big(\widehat{\mathrm{Var}}(\hat{t}_{1y}) + \widehat{\mathrm{Var}}\big(\hat{t}_{2y}^{(2)}\big) + 2\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}, \hat{t}_{2y}^{(2)}\big)\big), \tag{21}$$

$$\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}, \hat{t}_{2y}^{(2)}\big) = \sum_{k \in s_1}\sum_{l \in s_2} \frac{y_k}{\pi_{1k}} \frac{y_l}{\pi'_{2l}} \frac{\pi_{1k} - \pi_{1kl}^{(1)}}{\pi_{2kl}^{\star}} - \hat{t}_{1y}\hat{t}_{2y}^{(2)}, \tag{22}$$

$$\pi_{2kl}^{\star} = \mathbf{P}(k \in s_1, l \in s'_2)\mathbf{P}(k \in s_1, l \in s_2 \mid s'_2) = \pi_{1kl}\mathbf{P}(l \in s_2 \mid s'_2).$$

2. By a linear combination of $\hat{t}_{1y}^{\mathrm{rat}}$ and $\hat{t}_{2y}^{(2)}$, we obtain the estimator with the use of auxiliary information of the total

$$\hat{t}_2^{\mathrm{rat}} = \frac{1}{2}\big(\hat{t}_{1y}^{\mathrm{rat}} + \hat{t}_{2y}^{(2)}\big). \tag{23}$$

The expression for the variance of estimator (23):

$$\mathrm{Var}\big(\hat{t}_2^{\mathrm{rat}}\big) = \frac{1}{4}\big(\mathrm{Var}\big(\hat{t}_{1y}^{\mathrm{rat}}\big) + \mathrm{Var}\big(\hat{t}_{2y}^{(2)}\big) + 2\,\mathrm{Cov}\big(\hat{t}_{1y}^{\mathrm{rat}}, \hat{t}_{2y}^{(2)}\big)\big). \tag{24}$$

The variance $\mathrm{Var}(\hat{t}_2^{\mathrm{rat}})$ is estimated by

$$\widehat{\mathrm{Var}}\big(\hat{t}_2^{\mathrm{rat}}\big) = \frac{1}{4}\big(\widehat{\mathrm{Var}}\big(\hat{t}_{1y}^{\mathrm{rat}}\big) + \widehat{\mathrm{Var}}\big(\hat{t}_{2y}^{(2)}\big) + 2\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}^{\mathrm{rat}}, \hat{t}_{2y}^{(2)}\big)\big) \tag{25}$$

with the covariance estimator

$$\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}^{\mathrm{rat}}, \hat{t}_{2y}^{(2)}\big) = \sum_{k \in s_1}\sum_{l \in s_2} \frac{y_k - rx_k}{\pi_{1k}} \frac{y_k}{\pi'_{2l}} \frac{\pi_{1k} - \pi_{1kl}^{(1)}}{\pi_{2kl}^{\star}}.$$

3. By a linear combination of $\hat{t}_{1y}$, $\hat{t}_{2y}^{(2)}$, $\hat{t}_{3y}^{(3)}$ and $\hat{t}_{4y}^{(4)}$, we obtain a new estimator without the use of auxiliary information

$$\hat{t}_4 = \frac{1}{4}\big(\hat{t}_{1y} + \hat{t}_{2y}^{(2)} + \hat{t}_{3y}^{(3)} + \hat{t}_{4y}^{(4)}\big). \tag{26}$$

4. By a linear combination of $\hat{t}_{1y}^{\mathrm{rat}}$, $\hat{t}_{2y}^{\mathrm{rat}}$, $\hat{t}_{3y}^{\mathrm{rat}}$ and $\hat{t}_{4y}^{(4)}$ we obtain a new estimator of the total with the use of auxiliary information

$$\hat{t}_4^{\mathrm{rat}} = \frac{1}{4}\big(\hat{t}_{1y}^{\mathrm{rat}} + \hat{t}_{2y}^{\mathrm{rat}} + \hat{t}_{3y}^{\mathrm{rat}} + \hat{t}_{4y}^{(4)}\big). \tag{27}$$

Further, we are interested in the estimation of the finite population total t_y using two-phase and four-phase sampling schemes, when simple random samples of households without replacement and samples with probabilities proportional to household size without replacement are drawn in each of the phases.

5 Special cases of sampling design

5.1 Simple random sampling of households without replacement

5.1.1 Two-phase sampling scheme

Data for two quarters are used for the estimation of the population total t_y. Assume that s_1 of size n_1 is a simple random sample from the population $\mathcal{U}$, and its complement $s_2' = \mathcal{U} \setminus s_1$ of size $N - n_1$ is also a simple random sample from the population $\mathcal{U}$. s_2 of size n_2 is a simple random sample from s_2'. Then the first- and second-order inclusion probabilities to be used for (19) and (23) are calculated as follows:

$$\begin{aligned}
\pi_{1i} &= \mathbf{P}(i \in s_1) = \frac{n_1}{N}, \\
\pi_{1ij} &= \mathbf{P}(i \in s_1, j \in s_1) = \frac{n_1}{N}\frac{n_1 - 1}{N - 1}, \quad i \neq j, \\
\pi'_{2i} &= \mathbf{P}(i \in s_2') = \frac{N - n_1}{N}, \\
\pi'_{2ij} &= \mathbf{P}(i \in s_2', j \in s_2') = 1 - \frac{n_1}{N}\frac{n_1 - 1}{N - 1} - 2\frac{n_1}{N - 1}\left(1 - \frac{n_1}{N}\right), \quad i \neq j, \\
\pi_{2i|s_2'} &= \mathbf{P}(i \in s_2 \mid s_2') = \frac{n_2}{N - n_1}, \\
\pi_{2ij|s_2'} &= \mathbf{P}(i \in s_2, j \in s_2 \mid s_2') = \frac{n_2}{N - n_1}\frac{n_2 - 1}{N - n_1 - 1}, \quad i \neq j.
\end{aligned}$$

In the case of simple random sampling in each of the phases, the estimator of total (19) without the use of auxiliary information can be rewritten as

$$\hat{t}_2 = \frac{1}{2}\left(\frac{N}{n_1}\sum_{i \in s_1} y_i + \frac{N}{n_2}\sum_{i \in s_2} y_i\right). \tag{28}$$

The variance of the estimator $\hat{t}_2$ of the total t_y is expressed

$$\mathrm{Var}(\hat{t}_2) = \frac{1}{4}\left(N^2\left(1 - \frac{n_1}{N}\right)\frac{s_{1y}^2}{n_1} + N^2\left(1 - \frac{n_2}{N}\right)\frac{s_{2y}^2}{n_2} - 2Ns_{12y}^2\right), \tag{29}$$

where $s_{1y}^2 = s_{2y}^2 = s_{12y}^2 = (N-1)^{-1}\sum_{i=1}^N (y_i - \mu_y)^2$, $\mu_y = t_y/N$.

The variance $\mathrm{Var}(\hat{t}_2)$ is estimated unbiasedly by

$$\widehat{\mathrm{Var}}(\hat{t}_2) = \frac{1}{4}\left(N^2\left(1 - \frac{n_1}{N}\right)\frac{\hat{s}_{1y}^2}{n_1} + N^2\left(1 - \frac{n_2}{N}\right)\frac{\hat{s}_{2y}^2}{n_2} - 2N\hat{s}_{12y}^2\right), \tag{30}$$

where

$$\hat{s}_{1y}^2 = \frac{1}{n_1 - 1}\sum_{i \in s_1}\left(y_i - \frac{1}{n_1}\sum_{j \in s_1} y_j\right)^2, \quad \hat{s}_{2y}^2 = \frac{1}{n_2 - 1}\sum_{i \in s_2}\left(y_i - \frac{1}{n_2}\sum_{j \in s_2} y_j\right)^2,$$

$$\hat{s}^2_{12y} = \frac{1}{n_1+n_2-1}\sum_{i\in s}\biggl(y_i - \frac{1}{n}\sum_{j\in s} y_j\biggr)^2, \quad s = s_1 \cup s_2.$$

In the case of simple random sampling, in each of the phases, the estimator of total (23) with the use of auxiliary information can be rewritten as

$$\hat{t}_2^{\text{rat}} = \frac{1}{2}\biggl(t_{1x}^{(-1)}\frac{\sum_{i\in s_1} y_i}{\sum_{i\in s_1} x_i} + \frac{N}{n_2}\sum_{i\in s_2} y_i\biggr). \tag{31}$$

Here $t_{1x}^{(-1)}$ is the population total of the study variable y in the previous survey (-1 wave).

An approximate variance of the estimator $\hat{t}_2^{\text{rat}}$ of the total t_y is expressed

$$\text{AVar}\bigl(\hat{t}_2^{\text{rat}}\bigr) = \frac{1}{4}\biggl(N^2\Bigl(1-\frac{n_1}{N}\Bigr)\frac{s^2_{1\,y-rx}}{n_1} + N^2\Bigl(1-\frac{n_2}{N}\Bigr)\frac{s^2_{2y}}{n_2} - 2Ns^2_{12\,y-rx,y}\biggr), \tag{32}$$

where

$$s^2_{1\,y-rx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - rx_i)^2, \quad s^2_{2y} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu_y)^2,$$
$$s^2_{12\,y-rx,y} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - rx_i)(y_i-\mu_y), \quad r = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}, \quad \mu_y = \frac{t_y}{N}.$$

The variance $\text{Var}(\hat{t}_2^{\text{rat}})$ is estimated by

$$\widehat{\text{Var}}\bigl(\hat{t}_2^{\text{rat}}\bigr) = \frac{1}{4}\biggl(N^2\Bigl(1-\frac{n_1}{N}\Bigr)\frac{\hat{s}^2_{1\,y-rx}}{n_1} + N^2\Bigl(1-\frac{n_2}{N}\Bigr)\frac{\hat{s}^2_{2y}}{n_2} - 2N\hat{s}^2_{12\,y-rx,y}\biggr), \tag{33}$$

where

$$\hat{s}^2_{1\,y-rx} = \frac{1}{n_1-1}\sum_{i\in s_1}(y_i - \hat{r}x_i)^2, \quad \hat{r} = \frac{\sum_{j\in s_1} y_j}{\sum_{j\in s_1} x_j},$$
$$\hat{s}^2_{2y} = \frac{1}{n_2-1}\sum_{i\in s_2}\biggl(y_i - \frac{1}{n_2}\sum_{j\in s_2} y_j\biggr)^2,$$
$$\hat{s}^2_{12\,y-rx,y} = \frac{1}{n_1-1}\sum_{i\in s_1}(y_i-\hat{r}x_i)\biggl(y_i - \frac{1}{n_1}\sum_{j\in s_1} y_j\biggr).$$

The number of simple random sampling phases is further increased.

5.1.2 Four-phase sampling scheme

Data for four quarters are used for the estimation of the population total t_y. Assume that all samples in the selection procedure shown in Fig. 2 are simple random samples: s_1 of size n_1, $s_2' = \mathcal{U}\setminus s_1$ of size $N-n_1$, s_2 of size n_2, $s_3' = \mathcal{U}\setminus(s_1\cup s_2)$ of size $N-n_1-n_2$, s_3 of size n_3, $s_4' = \mathcal{U}\setminus(s_1\cup s_2\cup s_3)$ of size $N-n_1-n_2-n_3$, and s_4 of size n_4. The first- and second-order inclusion probabilities for samples s_1, s_2' and s_2 are introduced in Subsection 5.1.1. The corresponding first-order inclusion probabilities for samples s_3', s_3, s_4', s_4 to be used for (26) and (27) are calculated as follows:

$$\begin{aligned}
\pi'_{3i|s_2'} &= \mathbf{P}(i \in s_3' \mid s_2') = \frac{N - n_1 - n_2}{N - n_1},\\
\pi_{3i|s_3'} &= \mathbf{P}(i \in s_3 \mid s_3') = \frac{n_3}{N - n_1 - n_2},\\
\pi'_{4i|s_3'} &= \mathbf{P}(i \in s_4' \mid s_3') = \frac{N - n_1 - n_2 - n_3}{N - n_1 - n_2},\\
\pi_{4i|s_4'} &= \mathbf{P}(i \in s_4 \mid s_4') = \frac{n_4}{N - n_1 - n_2 - n_3}.
\end{aligned}$$

Now, in the case of simple random sampling, in each of the phases, estimator (26) without the use of auxiliary information of the total can be rewritten as

$$\hat{t}_4 = \frac{1}{4}\biggl(\frac{N}{n_1}\sum_{i\in s_1} y_i + \frac{N}{n_2}\sum_{i\in s_2} y_i + \frac{N}{n_3}\sum_{i\in s_3} y_i + \frac{N}{n_4}\sum_{i\in s_4} y_i\biggr). \tag{34}$$

In the case of simple random sampling, in each of the phases, estimator (27) with the use of auxiliary information of the total can be rewritten as

$$\hat{t}_4^{\mathrm{rat}} = \frac{1}{4}\biggl(t_{1x}^{(-1)}\frac{\sum_{i\in s_1} y_i}{\sum_{i\in s_1} x_i} + t_{2x}^{(-3)}\frac{\sum_{i\in s_2} y_i}{\sum_{i\in s_2} x_i} + t_{3x}^{(-1)}\frac{\sum_{i\in s_3} y_i}{\sum_{i\in s_3} x_i} + \frac{N}{n_4}\sum_{i\in s_4} y_i\biggr). \tag{35}$$

Remark 1. The totals $t_x^{(-1)}$, $t_{2x}^{(-3)}$, $t_{3x}^{(-1)}$ in (31), (35) cannot be known in a real survey, because the values of the study variable in all waves are known only for sample data. Therefore, we suggest replacing them by the estimates of t_y obtained for the corresponding wave using the whole sample consisting of four rotation parts, and to consider them further as fixed.

5.2 Unequal probability sampling of households without replacement with probabilities proportional to their size (successive sampling)

The selection of households through individuals is taken as an example of an unequal probability sampling design, which is used for the Lithuanian LFS. An individual is selected from a list with equal selection probabilities, and his/her household is included in the sample. Individual selection is repeated. If the household selected is already in the sample, this selection step is ignored. Otherwise, the household is included in the sample. The process is continued until the predetermined number n of different households is selected. This household selection scheme is studied by Rosén [20] and is called successive sampling design. The larger the household, the higher its probability to be selected for the sample, because any of its members can be selected from the list of individuals.

5.2.1 Order sampling designs

The order sampling design [18] is defined as follows. To each population element $i \in \mathcal{U}$, a probability distribution F_i is assigned, $i = 1, 2, \ldots, N$. Independent ranking random variables $Q_1, Q_2, \ldots, Q_N$ with distributions $F_1, F_2, \ldots, F_N$ are realized. The elements with the n smallest Q values constitute a sample. The distributions $F_1, F_2, \ldots, F_N$ are called ranking distributions.

Let us consider ranking distributions $F_i(u) = H(u; \lambda_i)$ with a shape distribution function $H(u)$, concentrated on a positive half-line, and real constants $\lambda_i > 0$, called intensities, desired or target inclusion probabilities, $i = 1, 2, \ldots, N$. The class of order sampling designs includes successive and Pareto sampling designs.

Successive sampling design has an exponential shape distribution function $H(u) = 1 - \mathrm{e}^{-u}$, $0 \leqslant u < \infty$. The ranking variables are

$$Q_i = \frac{\ln(1 - U_i)}{\ln(1 - \lambda_i)}, \quad i \in \mathcal{U}, \tag{36}$$

with the values of the random variables U_i distributed uniformly on $[0, 1]$.

The *Pareto sampling design* has a Pareto shape distribution function $H(u) = u/(1+u)$, $0 \leqslant u < \infty$. The ranking variables are

$$Q_i = \frac{U_i(1 - \lambda_i)}{\lambda_i(1 - U_i)}, \quad i \in \mathcal{U},$$

with the values of the random variables U_i distributed uniformly on $[0, 1]$.

For these sampling designs, the exact inclusion probabilities are approximately equal to the desired inclusion probabilities λ_i. According to Rosén [19], the inclusion probabilities for order sampling designs are asymptotically equal to the desired inclusion probabilities λ_i.

In our case, m_i is the number of household members, $M = \sum_{i=1}^{N} m_i$ is the total number of individuals, and we choose $\lambda_i = nm_i/M$. We obtain that with such a selection of desired inclusion probabilities, the class of order sampling designs intersects with the class of sampling designs with inclusion probabilities approximately proportional to the size measure (PPS).

The *conditional Poisson sampling* scheme (CP) is defined as follows: each unit in the population is selected with a prescribed probability p_i, $p_i > 0$, and $\sum_{i=1}^{N} p_i = n$, but only the samples of the desired size n are accepted. The inclusion probabilities for the CP sample will not be exactly equal to p_i, but only approximately.

Expressions for the second-order inclusion probabilities π_{1ij} in the case of order sampling design are presented in Aires [1], but they are too complex, and their computation is time-consuming.

According to Bondesson et al. [6, p. 700], Pareto and CP sampling designs are close for $p_i = \lambda_i$, and inclusion probabilities of all orders for Pareto sampling design may be approximated by the corresponding ones for the CP design. Relying on the approximate equality of the first-order inclusion probabilities, we will approximate the second-order inclusion probabilities for a successive design with the second-order inclusion probabilities for the CP design.

5.2.2 Two-phase sampling scheme

Data for two quarters are used for the estimation of the population total t_y. Assume that a sample s_1 of size n_1, a sample $s_2' = \mathcal{U} \setminus s_1$ of size $N - n_1$, and a sample s_2 of size n_2 are drawn according to the successive sampling design introduced by Rosén [18]. The first-phase household target inclusion probability in the sample s_1 is defined as $\lambda_i = n_1 m_i / M$, $i \in s_1$. The second-phase household target inclusion probability in the sample s_2 is defined as $\lambda_{2i} = n_2 m_i / \sum_{j \in s_2'} m_j$, $i \in s_2'$.

The first-order inclusion probabilities to be used for the estimation of the total t_y in (19) and (23) are expressed approximately as follows:

$$\pi_{1i} = \mathbf{P}(i \in s_1) \approx \lambda_i = \frac{n_1 m_i}{M}, \qquad \pi'_{2i} = \mathbf{P}(i \in s_2') \approx \lambda'_{2i} = \frac{M - n_1 m_i}{M},$$
$$\pi_{2i|s_2'} = \mathbf{P}(i \in s_2 \mid s_2') \approx \lambda_{2i} = \frac{n_2 m_i}{M_2}, \quad M_2 = \sum_{j \in s_2'} m_j.$$

The estimators $\hat{t}_{1y} = \hat{t}^{\lambda}_{1y} = \sum_{i \in s_1} y_i / \lambda_i$, $\hat{t}_{1x} = \hat{t}^{\lambda}_{1x} = \sum_{i \in s_1} x_i / \lambda_i$ are used in (19), (5) and (23).

For the estimators of variances (21), (25), the second-order inclusion probabilities for a successive sampling design are needed. Second-order inclusion probabilities for CP sampling $\breve{\pi}_{1ij}$, presented by Aires [1], are

$$\breve{\pi}_{1ij} = \frac{1}{\gamma_{1i} - \gamma_{1j}} (\gamma_{1i} \breve{\pi}_{1j} - \gamma_{1j} \breve{\pi}_{1i}) \quad \text{for } \gamma_{1i} \neq \gamma_{1j}, \tag{37}$$
$$\breve{\pi}_{1ij} = \frac{1}{k_{1i}} \biggl((n_1 - 1) \breve{\pi}_{1i} - \sum_{j\colon \gamma_{1j} \neq \gamma_{1i}} \breve{\pi}_{1ij} \biggr) \quad \text{for } \gamma_{1i} = \gamma_{1j}, \tag{38}$$

$i, j \in \mathcal{U}$, $i \neq j$. Here $\gamma_{1i} = p_{1i}/(1 - p_{1i})$, and k_{1i} is the number of elements with $j \in \mathcal{U}$, $j \neq i$, such that $\gamma_{1i} = \gamma_{1j}$; the probability p_{1i} is a selection probability of the element i for a CP sampling design, $i \in \mathcal{U}$. Suppose $\breve{\pi}_{1i}$ are given inclusion probabilities to the CP design equal to λ_i. Keeping them as known, we use the approximation result of Bondesson et al. [6, p. 705] to express p_{1i} through these inclusion probabilities:

$$\gamma_{1i} = \frac{p_{1i}}{1 - p_{1i}} \propto \frac{\breve{\pi}_{1i}}{1 - \breve{\pi}_{1i}} \exp\biggl\{ \frac{1/2 - \breve{\pi}_{1i}}{d} \biggr\}, \quad d = \sum_{i=1}^{N} \breve{\pi}_{1i} (1 - \breve{\pi}_{1i}).$$

Inserting the γ_{1i} obtained into (37), we find second-order inclusion probabilities to the CP design $\breve{\pi}_{1ij}$.

Approximating $P(i \in s_1, j \in s_1)$ in (8) by $\breve{\pi}_{1ij}$, we obtain the second-order inclusion probabilities for the sample s_2', and keeping $\pi_{2i|s_2'}$ as known, we obtain the second-order inclusion probabilities for s_2 ([1]) as follows:

$$\begin{aligned}
\pi'_{2ij} &= \mathbf{P}(i \in s'_2,\, j \in s'_2)\\
&\cong 1 - \check{\pi}_{1ij} - \frac{n_1 m_i}{M - m_j}(1-\lambda_j) - \frac{n_1 m_j}{M - m_i}(1-\lambda_i);\\
\check{\pi}_{2ij|s'_2} &= \frac{1}{\gamma_{2i} - \gamma_{2j}}\left(\gamma_{2i}\check{\pi}_{2j|s'_2} - \gamma_{2j}\check{\pi}_{2i|s'_2}\right) \quad \text{for } \gamma_{2i} \neq \gamma_{2j},\\
\check{\pi}_{2ij|s'_2} &= \frac{1}{k_{2i}}\left((n_2 - 1)\check{\pi}_{2i|s'_2} - \sum_{j:\,\gamma_{2j}\neq\gamma_{2i}} \check{\pi}_{2ij|s'_2}\right) \quad \text{for } \gamma_{2i} = \gamma_{2j}.
\end{aligned}$$

Here k_{2i} is the number of elements $j \neq i$ such that $\gamma_{2i} = \gamma_{2j}$ and

$$\gamma_{2i} = \frac{\lambda_{2i}}{1-\lambda_{2i}} \exp\left\{\frac{1/2 - \lambda_{2i}}{d}\right\}, \quad d = \sum_{i\in s'_2} \lambda_{2i}(1-\lambda_{2i}).$$

We have $\check{\pi}_{2ii|s'_2} = \check{\pi}_{2i|s'_2}$ in the case of $i = j$.

We remind that the second-order inclusion probabilities for a successive sampling design are approximated by the corresponding probabilities for the CP design.

After replacing the π values in (4), (7) and (12) with the corresponding approximate values presented in this section we estimate the variances of the estimators $\hat{t_2}$ and $\hat{t}_2^{\mathrm{rat}}$ of the total t_y in (21) and (25) respectively with

$$\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}, \hat{t}_{2y}^{(2)}\big) = \frac{M^2}{n_1 n_2}\Big(M - \sum_{i\in s_1} m_i\Big) \sum_{k\in s_1}\sum_{l\in s_2} \frac{y_k}{m_k}\frac{y_l}{m_l}\frac{1}{M - n_1 m_l} - \hat{t}_{1y}\hat{t}_{2y}^{(2)}, \tag{39}$$

$$\widehat{\mathrm{Cov}}\big(\hat{t}_{1y}^{\mathrm{rat}}, \hat{t}_{2y}^{(2)}\big) = \frac{M^2}{n_1 n_2}\Big(M - \sum_{i\in s_1} m_i\Big) \sum_{k\in s_1}\sum_{l\in s_2} \frac{y_k - \widehat{r}x_k}{m_k}\frac{y_l}{m_l}\frac{1}{M - n_1 m_l}, \tag{40}$$

$$\hat{r} = \frac{\hat{t}_{1y}}{\hat{t}_{1x}}, \quad \hat{t}_{1y} = \hat{t}_{1y}^{\lambda} = \sum_{i\in s_1}\frac{y_i}{\lambda_i} \sim \sum_{i\in s_1}\frac{y_i}{\pi_{1i}}, \quad \hat{t}_{1x} = \hat{t}_{1x}^{\lambda} = \sum_{i\in s_1}\frac{x_i}{\lambda_i} \sim \sum_{i\in s_1}\frac{x_i}{\pi_{1i}}.$$

The number of unequal probability sampling phases is further increased.

5.2.3 Four-phase sampling scheme

Data for four quarters are used for the estimation of the population total t_y. Assume that all samples in the selection procedure shown in Fig. 2 are drawn according to a successive sampling design: s_1 of size n_1, $s'_2 = \mathcal{U} \setminus s_1$ is of size $N - n_1$, s_2 is of size n_2, $s'_3 = \mathcal{U} \setminus (s_1 \cup s_2)$ of size $N - n_1 - n_2$, s_3 is of size n_3, $s'_4 = \mathcal{U} \setminus (s_1 \cup s_2 \cup s_3)$ is of size $N - n_1 - n_2 - n_3$ and s_4 is of size n_4. The first- and second-order inclusion probabilities for samples s_1, s'_2 and s_2 are introduced in Subsection 5.2.2. The corresponding first-order inclusion probabilities for samples s'_3, s_3, s'_4, s_4 for the sampling design under study, to be used for (26) and (27), are approximated as follows:

$$\begin{aligned}
\pi'_{3i|s'_2} &= \mathbf{P}(i \in s'_3 \mid s'_2) \cong \frac{M - n_1 m_i - n_2 m_i}{M - n_1 m_i},\\
\pi_{3i|s'_3} &= \mathbf{P}(i \in s_3 \mid s'_3) \cong \frac{n_3 m_i}{M_3}, \quad M_3 = \sum_{j \in s'_3} m_j,\\
\pi'_{4i|s'_3} &= \mathbf{P}(i \in s'_4 \mid s'_3) \cong \frac{M - n_1 m_i - n_2 m_i - n_3 m_i}{M - n_1 m_i - n_2 m_i},\\
\pi_{4i|s'_4} &= \mathbf{P}(i \in s_4 \mid s'_4) \cong \frac{n_4 m_i}{M_4}, \quad M_4 = \sum_{j \in s'_4} m_j.
\end{aligned}$$

Second-order inclusion probabilities for a four-phase sampling design are not presented here. Only the empirical variance of the estimates in the case of a four-phase sampling scheme is used in the simulation study.

6 Simulation study

In this section, we present a simulation study for the comparison of the performance of several estimators of the total using data of two and four quarters, with simple random sampling (SRS) and sampling with probability proportional to size (PPS) (successive order sampling) of households without replacement in each of the phases.

We study the real LFS data of Statistics Lithuania. The study population consists of $N = 500$ households. The variables of interest, y and x, are the number of employed (or unemployed) individuals in the population of households in the current and previous waves. The population totals $t_{1x}^{(-1)}$, $t_{2x}^{(-3)}$, $t_{3x}^{(-1)}$ are available and are used for estimators (23), (27). The correlation coefficient between the variables x and y in the household population for the number of employed individuals of interest is $\rho(x, y) = 0.95$. It means a strong linear relationship. For the number of unemployed individuals of interest, the correlation coefficient is $\rho(x, y) = 0.80$.

For the two-phase sampling scheme, $B = 500$ samples s_1 and s_2 of size $n_1 = n_2 = 100$ ($n = n_1 + n_2 = 200$) are selected by simple random sampling and successive sampling.

For the four-phase sampling scheme, samples s_1, s_2, s_3 and s_4 of size $n_1 = n_2 = n_3 = n_4 = 50$ ($n = n_1 + n_2 + n_3 + n_4 = 200$) are selected by simple random sampling and successive sampling.

For each of the estimators $\hat{t}_2$, $\hat{t}_2^{\mathrm{rat}}$, $\hat{t}_4$ and $\hat{t}_4^{\mathrm{rat}}$, we have calculated the estimates of the population total t_y of the study variable y. For all estimators $\hat{\theta} = \hat{t}_2, \hat{t}_2^{\mathrm{rat}}, \hat{t}_4, \hat{t}_4^{\mathrm{rat}}$, the averages of the estimates, the averages of the variance estimates and the empirical variances of the estimates

$$\bar{\hat{\theta}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i, \qquad \overline{\widehat{\mathrm{Var}}}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}\widehat{\mathrm{Var}}(\hat{\theta}_i), \qquad \mathrm{Var}_{\mathrm{emp}}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\hat{\theta}})^2$$

Table 1. Average of variance estimates and empirical variances for each part of the combined estimator in two-phase successive sampling.

Estimator		Number of employed		Number of unemployed	
		Empirical	Aver. estimate	Empirical	Aver. estimate
$\hat{t}_{1y}^{\mathrm{HT}}$	1st part	1 177	1 331	209	197
$\hat{t}_{1y}^{\mathrm{rat}}$	1st part rat.	208	180	120	83
$\hat{t}_{2y}^{(2)}$	2nd part	1 320	1 296	195	198
$\hat{t}_2$	Combined	490	657	74	99
$\hat{t}_2^{\mathrm{rat}}$	Comb. rat.	363	369	76	70

are calculated. The results of the simulation are presented in Table 1. They illustrate the averages of the estimates of variances and empirical variances for each part of the combined estimators $\hat{t}_2$ and $\hat{t}_2^{\mathrm{rat}}$, when samples are drawn according to the successive sampling design in each phase.

The combined estimator has a smaller variance than any of its parts first of all because of a higher sample size used to calculate it. Using the ratio estimator, the empirical variance of the estimator of the total decreases for the number of employed individuals and has small effect on the estimates of the number of unemployed individuals.

In almost all cases, the calculated averages of the estimates of variances are close to the empirical variances of the estimates for all parts of combined estimators $\hat{t}_2$ and $\hat{t}_2^{\mathrm{rat}}$ using a two-phase sampling design. It shows that the expression of variance of a combined estimator of the total obtained for two-phase successive sampling is accurate enough, and the inclusion probabilities for a successive sampling design can be approximated by the corresponding probabilities of unequal probability without replacement conditional Poisson sampling design to calculate estimates of the proposed estimators of the totals and their variance estimators.

The box-plot diagrams of the estimates of the number of employed and unemployed individuals in the household population using the proposed estimators $\hat{t}_2$, $\hat{t}_2^{\mathrm{rat}}$, $\hat{t}_4$ and $\hat{t}_4^{\mathrm{rat}}$ are presented in Figs. 3 and 4. Simple random samples (SRS) and successive samples with probabilities proportional to the household size (PPS) are drawn in each of the phases using two-phase and four-phase sampling schemes.

The estimates of the number of employed individuals have a lower variance using the PPS sampling design, compared to the SRS sampling design. The estimates calculated for the combined ratio estimator of the total for the four-phase sampling scheme has lower variance than the estimates calculated for the combined estimator of the total without the use of auxiliary information. The estimates with the lowest variance are obtained by the combined ratio estimator of the total for four-phase sampling with PPS sampling in each of the phases. The variances of the estimates of the number of unemployed individuals do not differ much.

The box-plot diagrams of the variance estimates of the number of employed and unemployed persons in the household population using all the estimators obtained $\hat{t}_2$, $\hat{t}_2^{\mathrm{rat}}$, $\hat{t}_4$ and $\hat{t}_4^{\mathrm{rat}}$ are presented in Fig. 5. The estimates of the variances of estimators of the number of employed individuals using the two-phase sampling scheme with PPS in each of the phases are lower than those obtained with the SRS sampling design. The

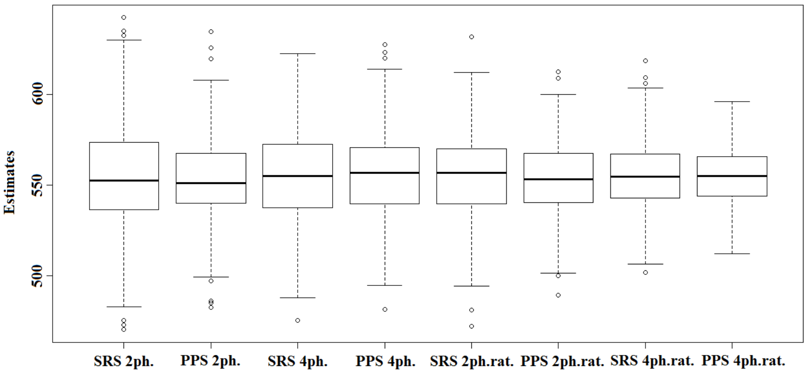


Figure 3. Estimates of the number of employed individuals.

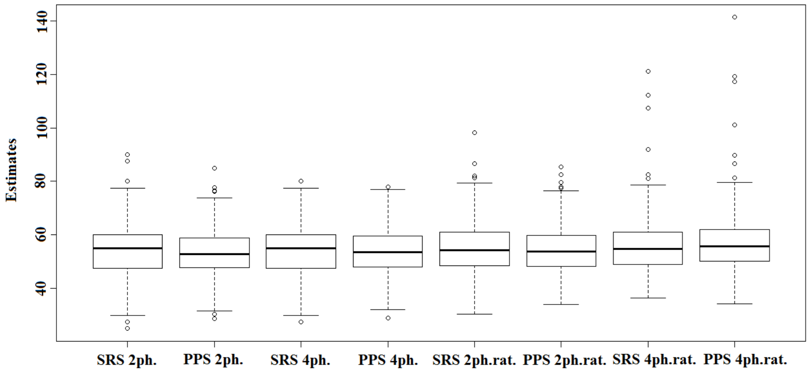


Figure 4. Estimates of the number of unemployed individuals.

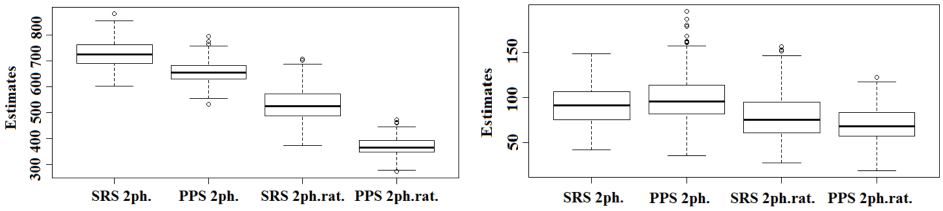


Figure 5. Estimates of the variances of estimators for the number of employed (left) and unemployed (right) individuals

lowest variance has a combined ratio estimator of the total with the use of auxiliary information. The variation of estimates for the variances of estimators of the number of unemployed individuals differs little, and this is because the correlation coefficient between the variables x and y in the household population for the number of unemployed individuals is lower than for the number of employed individuals.

Table 2. True variances, empirical variances, averages of variance estimates, and relative biases.

		Number of employed				Number of unemployed			
Estimator		Var. true	Var. emp.	Aver. of estimates	Relat. bias	Var. true	Var. emp.	Aver. of estimates	Relat. bias
$\hat{t}_2$	SRS	728	770	729	0.001	90	90	91	0.004
$\hat{t}_2^{\mathrm{rat}}$	SRS	525	507	530	0.003	77	96	79	0.018
$\hat{t}_2$	PPS	–	490	657	-0.003	–	74	99	-0.007
$\hat{t}_2^{\mathrm{rat}}$	PPS	–	363	369	0.000	–	76	70	0.004
$\hat{t}_4$	SRS	–	623	–	0.001	–	86	–	-0.004
$\hat{t}_4^{\mathrm{rat}}$	SRS	–	320	–	0.002	–	101	–	0.025
$\hat{t}_4$	PPS	–	534	–	0.003	–	75	–	-0.002
$\hat{t}_4^{\mathrm{rat}}$	PPS	–	238	–	0.002	–	109	–	0.049

Table 2 illustrates the true variances, averages of the estimates of variances, empirical variances and relative empirical biases:

$$\mathrm{RB}(\hat{\theta}) = \frac{1}{t_y}\frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i - t_y)$$

of each of the combined estimators obtained before, when SRS and PPS without replacement were drawn in each of the phases.

At present, only empirical variances of estimates have been calculated for the four-phase sampling schemes in the case of SRS and successive sampling, in each of the phases. Empirical variances of the estimates of the number of employed persons are lower in the case of a successive sampling design than in the case of SRS. The smallest empirical variance of the estimates of the number of employed individuals has been obtained for the combined ratio estimator when the four-phase sampling has been used. Relative biases for the estimator of employed individuals are insignificant. Combined ratio estimators for the number of the unemployed have a high variance and high relative bias in comparison with the estimates without auxiliary information.

Conclusions

The results of the simulation study show that:

- The ratio type combined estimator in a four-phase LFS sampling design has a lower variance for the estimate of the number of employed individuals and a higher variance for the estimates of the number of the unemployed in comparison with the corresponding estimators, which do not use auxiliary data.
- The approximation of the second order inclusion probabilities for a successive sampling with the corresponding probabilities for conditional Poisson sampling may be used to estimate the variances of the estimator in successive sampling.
- The successive sampling design used in the Lithuanian LFS is effective in the estimation of the number of employed individuals and its effectiveness is the same or even worse than for SRS in the estimation of the number of the unemployed.

7 Discussion

A two-phase sampling design with second-phase stratification by the household size has been used to estimate the number of employed and unemployed individuals in [14]. The simulation results show that the variance for the estimates of the number of employed individuals decreases significantly in comparison with the one-phase sampling design of the same size, and it does not decrease for the estimates of the number of the unemployed. As we see, the result of the [14] study leads to a similar conclusion as in the case of the current paper.

The combined ratio-type estimator may be effectively used in practice for the estimation of the number of employed individuals. When using ratio-type estimators, the data of the elements belonging to the current sample and to the sample of the previous wave, the data of the previous wave are needed. In the case of non-availability of the data of the previous wave for some elements, the values of the variables needed have to be imputed.

The ratio estimator used here is the simplest way to use auxiliary information at the estimation stage. A regression estimator of the total with the study variable of the previous wave as an auxiliary variable may also be used. A larger number of auxiliary variables from the previous waves and a calibrated estimator of the total instead of a ratio estimator is a possible generalization of the problem.

Acknowledgment. The authors are thankful to three anonymous referees for valuable comments, resulting in a significant improvement of the paper.

References

1. N. Aires, Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs, *Methodol. Comput. Appl. Probab.*, **1**(4):457–469, 1999.

2. A. Andersson, K. Andersson, P. Lundquist, Estimation of change in a rotation panel design, in *Bulletin of the International Statistical Institute Proceedings of the 58th World Statistics Congress, Dublin, August 21–26, 2011*, ISI, The Hague, 2012, pp. 4520–4525, http://2011.isiproceedings.org/papers/950903.pdf.

3. R. Arnab, Sampling on two occasions: Estimator of population total, *Survey Methodology*, **24**(2):185–192, 1998.

4. E. Artes, A. Garcia, Estimation of current population ratio in successive sampling, *J. Indian Soc. Agric. Stat.*, **54**(3):342–354, 2001.

5. Y.G. Berger, R. Priam, *A simple variance estimator of change for rotating repeated surveys: An application to the EU-SILC household surveys*, Southampton Statistical Sciences Research Institute, 2013, http://eprints.soton.ac.uk/347142/.

6. L. Bondesson, I. Traat, A. Lundqvist, Pareto sampling versus Sampford and conditional Poisson sampling, *Scand. J. Stat.*, **33**:699–720, 2006.

7. V. Chadyšas, D. Krapavickaitė, Estimation of employed persons in the case of sample rotation, *Liet. matem. rink. LMD darbai*, **48–49**:288–293, 2008 (in Lithuanian).

8. L. Fattorini, M. Marcheselli, C. Pisani, A three-phase sampling strategy for large-scale multiresource forest inventories, *J. Agric. Biol. Environ. Stat.*, **11**(3):296–316, 2006, doi:10.1198/108571106X130548.

9. W.A. Fuller, Estimation for multiple phase samples, in: R.L. Chambers, J. Skinner (Eds.), *Analysis of Survey Data*, John Wiley & Sons, Chichester, 2003, pp. 307–322.

10. N. Hamad, M. Hanif, N. Haider, A regression type estimator with two auxiliary variables for two-phase sampling, *Open Journal of Statistics*, **3**:74–78, 2013, `http://www.scirp.org/journal/ojs/`.

11. M.A. Hidiriglou, V. Estevao, Dealing with nonresponse using follow up, in *Proceedings of the Joint Statistical Meeting, Montréal, Québec, Canada, August 3–8, 2013*, American Statistical Association, 2013, pp. 1478–1489.

12. D.G. Horvitz, D.J. Thompson, A generalization of sampling without replacement from a finite universe, *J. Am. Stat. Assoc.*, **47**:663–685, 1952.

13. S. Jeyaratnam, D.C. Bowden, F.A. Graybill, W.E. Frayer, Estimation in multiphase designs for stratification, *Forest Sci.*, **30**(2):484–491, 1984.

14. D. Krapavickaitė, The first phase order sampling for the second phase stratification, in *Proceedings of the 59th World Statistics Congress of the International Statistical Institute, Hong Kong, August 25–30, 2013*, ISI, The Hague, 2013, pp. 3714–3719, `http://2013.isiproceedings.org/Files/CPS016-P8-S.pdf`.

15. R.D. Narain, On sampling without replacement with varying probabilities, *J. Indian Soc. Agric. Stat.*, **3**:169–174, 1951.

16. F.C. Okafor, H. Lee, Double sampling for ratio and regression estimation with sub-sampling the non-respondents, *Survey Methodology*, **26**(2):183–188, 2000.

17. L. Qualité, Y. Tillé, Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added, *Survey Methodology*, **34**(2):173–181, 2008.

18. B. Rosén, On sampling with probability proportional to size, R&D Report 1996:1, Statistics Sweden, 1996.

19. B. Rosén, Order πps inclusion probabilities are asymptotically correct, R&D Report 2001:2, Statistics Sweden, 2001.

20. B. Rosén, Variance estimation for systematic pps-sampling, R&D Report 1991:15, Statistics Sweden, 1991.

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

22. S. Singh, *Advanced Sampling Theory with Applications: How Michael Selected Amy*, Vols. 1,2, Kluwer Academic, The Netherlands, 2003.