INFORMATICA, 2005, Vol. 16, No. 3, 347–364

Estimation of a Ratio in the Finite Population

Danutė KRAPAVICKAITĖ, Aleksandras PLIKUSAS
Institute of Mathematics and Informatics
Akademijos 4, 08663 Vilnius, Lithuania
e-mail: {krapav,plikusas}@ktl.mii.lt

Received: November 2004

Abstract. Estimation of the ratio of two totals is considered, when a probability sample from the finite population is available. Four estimators of the ratio are examined. The first one – called "simple" – is the ratio of the Horvitz–Thompson estimators of totals; the second – the ratio of two ratio estimators of totals; the third one – the ratio of two regression estimators of totals. The fourth one is a calibrated estimator of the ratio. The variances of these estimators are compared. The properties of such estimators of the ratio are studied. The simulation results are presented.

Key words: finite population, probability sample, ratio of two totals, calibrated estimator.

1. Introduction

Consider a finite population $\mathcal{U} = \{1, 2, \ldots, N\}$. Let y and z be two study variables defined on the population $\mathcal{U}$ and taking values $\{y_1, \ldots, y_N\}$ and $\{z_1, \ldots, z_N\}$ respectively. The values of the variables y and z are not known. Denote the unknown population totals of these variables by

$$t_y = \sum_{k=1}^{N} y_k, \quad t_z = \sum_{k=1}^{N} z_k.$$

We are interested in the estimation of the ratio of two totals

$$R = t_y / t_z \,.$$

Such a parameter is often met in the official statistics. For example, the unemployment rate equals the ratio between the number of unemployed individuals and the number of individuals in the labor force in the country. The average salary equals the ratio of the sum of salary funds of enterprises to the sum of the number of employees. Income per capita is equal to the sum of income of households divided by the sum of the numbers of their members.

The unknown parameter of interest is estimated from the sample. To this end probability samples are used. It means that the sampling method used satisfies the following conditions: the set of all possible samples is defined, the probability of each sample which

can be selected is pointed out and the sum of these probabilities equals 1; it is required that each element of the survey population has a positive probability to be included into the sample. If it is possible to construct a sample selection mechanism satisfying the conditions mentioned above, the sample drawn using this mechanism is called a probability sample (Cochran, 1977; Särndal *et al.*, 1992).

The simplest and usual way to estimate a ratio is to use the Horvitz–Thompson estimator of totals (Särndal *et al.*, 1992) taking into account sampling design in the numerator and denominator of the ratio. The quality of the estimator can be improved by using known auxiliary variables. A number of different estimators can be constructed exploiting these additional variables.

The aim of this paper is, firstly, to propose some estimators of the ratio where distinct cases of the calibrated estimators of totals are used in the numerator and denominator: the ratio and regression estimators of totals; secondly, some properties of the calibrated estimator of the ratio, suggested by Plikusas (2001) are investigated. All the estimation methods are compared to some extent. The simulation study illustrates the theoretical results.

2. Estimation Methods of the Ratio

Denote by s, $s \subset \mathcal{U}$ a probability sample drawn from the finite population $\mathcal{U}$, $\pi_k = \mathbf{P}(k \in s)$ – a probability of any element $k \in \mathcal{U}$ to be included into the sample s, π_{kl} – the inclusion probability of a pair of elements (k, l) into the sample. The notation $\pi_{kk} = \pi_k$ is used for convenience. The numbers $d_k = 1/\pi_k$ are called the design weights. The values of all variables of the sampled elements are supposed to be known.

2.1. *Horvitz–Thompson Estimator of Totals*

The simplest and straightforward way to estimate the ratio R is to use the Horvitz–Thompson estimator for estimating totals and to take the ratio of these estimators as the estimator of the ratio.

In order to make the paper more self contained we will present the known result on the main properties of the Horvitz–Thompson estimator of a total.

PROPOSITION 1. The Horvitz–Thompson estimator of the total t_y

$$\widehat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

is unbiased. The variance of this estimator equals

$$Var\,(\widehat{t}_y) = \sum_{k,l \in \mathcal{U}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

The estimator of this variance

$$\widehat{Var}\,(\widehat{t}_y) = \sum_{k,l\in s} \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{1}$$

is unbiased, if $\pi_{kl} > 0$ for all pairs (k, l), $k, l \in \mathcal{U}$.

Proof. Let us introduce the sample membership indicators

$$I_k = \begin{cases} 1, & \text{if } k \in s, \\ 0, & \text{if } k \notin s, \end{cases}$$

$k = 1, \ldots, N$. Then the expectation $\mathbf{E}I_k = \pi_k$, $\mathbf{E}I_kI_l = \pi_{kl}$, the variance $Var\,(I_k) = \pi_k - \pi_k^2$ and the covariance $Cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$. Now we can calculate the expectation and variance of $\widehat{t}_y$:

$$\begin{aligned}
&\mathbf{E}\,\widehat{t}_y = \mathbf{E}\sum_{k\in s}\frac{y_k}{\pi_k} = \mathbf{E}\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}I_k = \sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}\,\mathbf{E}I_k = t;\\
&Var\,(\widehat{t}_y) = Var\left(\sum_{k\in\mathcal{U}}\frac{y_k}{\pi_k}I_k\right) = \sum_{k\in\mathcal{U}}\frac{y_k^2}{\pi_k^2}\,Var\,(I_k) + \sum_{\substack{k,l\in\mathcal{U}\\ k\neq l}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}\,Cov(I_k, I_l)\\
&\quad = \sum_{k,l\in\mathcal{U}}(\pi_{kl} - \pi_k\pi_l)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.
\end{aligned}$$

The unbiasedness of the variance estimator follows from the fact that

$$\mathbf{E}\,I_kI_l\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} = \pi_{kl} - \pi_k\pi_l.$$

Let $\widehat{t}_z = \sum_{k\in s} d_kz_k$ be a Horvitz–Thompson estimator of the total t_z. Let the simple estimator of the ratio $R = t_y/t_z$ be

$$\widehat{R} = \widehat{t}_y/\widehat{t}_z\,.$$

Using the first-order terms of the Taylor series expansion of this estimator, the approximate variance can be expressed as

$$AVar\,(\widehat{R}) = \frac{1}{t_z^2}Var\,(\widehat{t}_y - R\widehat{t}_z) = \frac{1}{t_z^2}Var\left(\sum_{k\in s}\frac{y_k - Rz_k}{\pi_k}\right). \tag{2}$$

We see that it can be calculated and estimated as the variance of the Horvitz–Thompson estimator of the total of the variable $y - Rz$. This linearization approach will be also applied for the other estimators considered in the paper.

The estimator of the variance of the estimator of ratio $\widehat{R}$ is constructed as follows. The quantity $1/t_z^2$ (see (2)) is estimated by $1/\widehat{t}_z^2$, and Horvitz–Thompson estimator (1) of the

variance of the total is used for the variable $y - Rz$, where R is estimated by $\widehat{R}$. So, we get

$$\widehat{Var}\,(\widehat{R}) = \frac{1}{\widehat{t}_z^2} \sum_{k,l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{v_k}{\pi_k} \frac{v_l}{\pi_l}, \quad v_k = y_k - \widehat{R} z_k. \tag{3}$$

Such type of variance estimators is commonly used when estimating the variances of a rational functions of the Horvitz–Thompson estimators of totals.

2.2. *Estimation of Ratio Using Ratio Estimators of Totals*

Let us suppose some auxiliary information is available. This auxiliary information may be known from the previous complete surveys of the same population, administrative registers and other sources. Let a variable x_y with the population values $x_{y1}, \ldots, x_{yN}$ and a variable x_z with the values $x_{z1}, \ldots, x_{zN}$ be auxiliary variables with the totals

$$t_{xy} = \sum_{k=1}^{N} x_{yk} = N\mu_{xy}, \quad t_{xz} = \sum_{k=1}^{N} x_{zk} = N\mu_{xz}.$$

The ratio and regression estimators of totals (Cochran, 1977) are constructed using the known auxiliary variables and may have smaller variances, provided that the study and auxiliary variables are correlated enough. So, naturally there arises an idea to use this kind of estimators for the totals in the numerator and denominator when estimating the ratio.

If the variable x_y is auxiliary for the variable y with the known population total t_{xy}, then the estimator

$$\widehat{t}_y^{(rat)} = \frac{\widehat{t}_y}{\widehat{t}_{xy}} t_{xy}, \quad \widehat{t}_{xy} = \sum_{k \in s} \frac{x_{yk}}{\pi_k},$$

is called a ratio estimator of the total t_y. The approximate variance of $\widehat{t}_y^{(rat)}$ is

$$\begin{aligned} AVar\left(\widehat{t}_y^{(rat)}\right) &= Var\left(\widehat{t}_y - \frac{t_y}{t_{xy}} \widehat{t}_{xy}\right) \\ &= Var\,(\widehat{t}_y) + \left(\frac{t_y}{t_{xy}}\right)^2 Var\,(\widehat{t}_{xy}) - 2\frac{t_y}{t_{xy}}\, Cov(\widehat{t}_y, \widehat{t}_{xy}). \end{aligned} \tag{4}$$

Denote by $cv(\widehat{\theta})$ the coefficient of variation of a random variable $\widehat{\theta}$:

$$cv(\widehat{\theta}) = \frac{\sqrt{Var(\widehat{\theta})}}{\mathbf{E}\,\widehat{\theta}}.$$

It follows from (4), that the approximate variance of the ratio estimator of total is not bigger than the variance of the Horvitz–Thompson estimator of the total $\widehat{t}_y$:

$$AVar\left(\widehat{t}_y^{(rat)}\right) \leqslant Var\left(\widehat{t}_y\right),$$

provided the coefficient of correlation ρ between $\widehat{t}_y$ and $\widehat{t}_{xy}$ is sufficiently high:

$$\rho\big(\widehat{t}_y, \widehat{t}_{xy}\big) \geqslant \frac{1}{2}\frac{cv(\widehat{t}_{xy})}{cv(\widehat{t}_y)}, \quad \rho\big(\widehat{t}_y, \widehat{t}_{xy}\big) = \frac{Cov(\widehat{t}_y, \widehat{t}_{xy})}{\sqrt{Var(\widehat{t}_y)}\sqrt{Var(\widehat{t}_{xy})}}. \tag{5}$$

Denote by $cv(y)$ the coefficient of variation of the variable y:

$$cv(y) = \frac{s_y}{\mu_y}, \quad s_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu_y)^2, \quad \mu_y = t_y/N.$$

In the case of simple random sampling (when elements are sampled with equal selection probabilities without replacement (see Sect. 3)) the condition (5) can be written as

$$\rho(y, x_y) \geqslant \frac{1}{2}\frac{cv(x_y)}{cv(y)}, \quad \rho(y, x_y) = \frac{\sum_{k\in\mathcal{U}}(y_k - \mu_y)(x_{yk} - \mu_{xy})}{\sqrt{\sum_{k\in\mathcal{U}}(y_k - \mu_y)^2}\sqrt{\sum_{k\in\mathcal{U}}(x_{yk} - \mu_{xy})^2}}.$$

Taking the ratio estimators of totals t_y and t_z, we obtain the estimator of the ratio R

$$\widehat{R}^{(rat)} = \frac{\frac{\widehat{t}_y}{\widehat{t}_{xy}}t_{xy}}{\frac{\widehat{t}_z}{\widehat{t}_{xz}}t_{xz}} = \frac{t_{xy}}{t_{xz}}\cdot\frac{\widehat{t}_y\widehat{t}_{xz}}{\widehat{t}_z\widehat{t}_{xy}} = R_0\,\frac{\widehat{t}_y\widehat{t}_{xz}}{\widehat{t}_z\widehat{t}_{xy}}, \quad \widehat{t}_{xz} = \sum_{k\in s}\frac{x_{zk}}{\pi_k}.$$

Here the ratio $R_0 = t_{xy}/t_{xz}$ is supposed to be known.

For an approximate expression of the variance of $\widehat{R}^{(rat)}$, the Taylor series expansion of this estimator by the first-order terms is used. Taking partial derivatives of $\widehat{R}^{(rat)}$ by $\widehat{t}_y$, $\widehat{t}_z$, $\widehat{t}_{xy}$, $\widehat{t}_{xz}$ at the point $(\widehat{t}_y, \widehat{t}_z, \widehat{t}_{xy}, \widehat{t}_{xz}) = (t_y, t_z, t_{xy}, t_{xz})$, we derive the linearized estimator

$$\widehat{R}_l^{(rat)} = R\Big(1 + \frac{\widehat{t}_y - R\widehat{t}_z}{t_y} - \frac{\widehat{t}_{xy} - R_0\widehat{t}_{xz}}{t_{xy}}\Big).$$

We consider its variance as the approximate variance of $\widehat{R}^{(rat)}$:

$$AVar(\widehat{R}^{(rat)}) = Var(\widehat{R}_l^{(rat)}) = \frac{1}{t_z^2}Var\big((\widehat{t}_y - R\widehat{t}_z) - R_{1\,rat}(\widehat{t}_{xy} - R_0\widehat{t}_{xz})\big).$$

Here the notation $R_{1\,rat} = t_y/t_{xy}$ is used. The expression (3) with

$$v_k = y_k - \widehat{R}z_k - \frac{\widehat{t}_y}{\widehat{t}_{xy}}x_{yk} + \frac{\widehat{t}_y}{\widehat{t}_{xz}}x_{zk}$$

is used as the estimator of variance of $\widehat{R}^{(rat)}$.

2.3. *Estimation of Ratio Using Regression Estimators of Totals*

Let the total t_{xy} of the auxiliary variable x_y be known. The estimator of the total t_y

$$\widehat{t}_y^{(reg)} = \widehat{t}_y + (t_{xy} - \widehat{t}_{xy})\widehat{B}_y$$

with

$$\widehat{B}_y = \frac{\sum_{k\in s} d_k(y_k - \widehat{\mu}_y)(x_{yk} - \widehat{\mu}_{xy})}{\sum_{k\in s} d_k(x_{yk} - \widehat{\mu}_{xy})^2} = \frac{\widehat{Cov}(y, x_y)}{\widehat{Var}(x_y)}, \tag{6}$$

$$\widehat{\mu}_y = \widehat{t}_y/N, \quad \widehat{\mu}_{xy} = \widehat{t}_{xy}/N,$$

is called the regression estimator of total (Cochran, 1977; Särndal *et al.*, 1992). Its approximate variance for any sample design is

$$AVar\left(\widehat{t}_y^{(reg)}\right) = \sum_{k,l\in\mathcal{U}} (\pi_{kl} - \pi_k\pi_l)\frac{y_k - A - B_y x_{yk}}{\pi_k}\frac{y_l - A - B_y x_{yl}}{\pi_l}.$$

Here

$$B_y = \frac{Cov(y, x_y)}{Var(x_y)} = \frac{\sum_{k=1}^N (y_k - \mu_k)(x_{yk} - \mu_{xy})}{\sum_{k=1}^N (x_{yk} - \mu_{xy})^2}, \quad A = \mu_y - B_y\mu_{xy}\,. \tag{7}$$

It is known that if the correlation coefficient $\rho(y, x_y) > 0$, then, in the case of simple random sampling, the approximate variance of the regression estimator is less than that of Horvitz–Thompson estimator:

$$AVar\left(\widehat{t}_y^{(reg)}\right) < Var\left(\widehat{t}_y\right).$$

Using this estimator in the numerator and denominator of the ratio $R = t_y/t_z$, we get the regression estimator of the ratio

$$\widehat{R}^{(reg)} = \frac{\widehat{t}_y + (t_{xy} - \widehat{t}_{xy})\widehat{B}_y}{\widehat{t}_z + (t_{xz} - \widehat{t}_{xz})\widehat{B}_z}$$

with $\widehat{B}_y$ defined in (6) and

$$\widehat{B}_z = \frac{\widehat{Cov}\,(z, x_z)}{\widehat{Var}\,(x_z)}.$$

The total of the variable x_z is also supposed to be known.

In order to get the expression of the approximate variance of the estimator $\widehat{R}^{(reg)}$, the expansion of this estimator in Taylor series up to the first-order terms is used. So, we get the linear estimator

$$\widehat{R}_l^{(reg)} = R + \frac{1}{t_z}\big((\widehat{t}_y - t_y) - B_y(\widehat{t}_{xy} - t_{xy}) - R(\widehat{t}_z - t_z) + RB_z(\widehat{t}_{xz} - t_{xz})\big),$$

where B_y is defined in (7),

$$B_z = \frac{\sum_{k=1}^{N}(z_k - \mu_z)(x_{zk} - \mu_{xz})}{\sum_{k=1}^{N}(x_{zk} - \mu_{xz})^2}, \quad \mu_z = t_z/N, \quad \mu_{xz} = t_{xz}/N.$$

Its variance is considered as the approximate variance of $\widehat{R}^{(reg)}$:

$$AVar\,(\widehat{R}^{(reg)}) = Var\,(\widehat{R}_l^{(reg)}) = \frac{1}{t_z^2} Var\,\big((\widehat{t}_y - R\widehat{t}_z) - (\widehat{t}_{xy}B_y - R\widehat{t}_{xz}B_z)\big).$$

The expression (3) with

$$v_k = y_k - \widehat{R}z_k - \widehat{B}_y x_{yk} + \widehat{R}\widehat{B}_z x_{zk}$$

is used as the estimator of variance of $\widehat{R}^{(reg)}$.

2.4. *Calibrated Estimator of the Ratio*

The calibrated estimator of the population total was introduced by Deville and Särndal (1992). In order to get a calibrated estimator of the total $t_y = y_1 + \ldots + y_N$, we are looking for the weights w_k, which differ as little as possible from the design weights d_k with arbitrary positive q_k in the sense that

$$\sum_{k\in s} \frac{(w_k - d_k)^2}{q_k d_k} \to \min, \tag{8}$$

and which satisfy the calibration equation

$$\sum_{k\in s} w_k x_{yk} = t_{xy}$$

for some auxiliary variable x_y with the known population total t_{xy}. The weights w_k are called calibrated weights, and the estimator

$$\widehat{t}_y^{(cal)} = \sum_{k\in s} w_k y_k$$

is called a calibrated estimator of the total t_y. The ratio and regression estimators of total are distinct separate cases of the calibrated estimator of total.

A calibrated estimator of a ratio is proposed by Plikusas (2001), where the ratio $R_0 = t_{xy}/t_{xz}$ of totals of the auxiliary variables x_y and x_z is kept fixed.

DEFINITION. The estimator of the ratio R

$$\widehat{R}_w = \frac{\sum_{k\in s}\, w_k\, y_k}{\sum_{k\in s}\, w_k\, z_k}$$

is called calibrated if the new weights w_k minimize the distance function (8) and estimate the known ratio R_0 of totals of auxiliary variables x_y and x_z without error:

$$\frac{\sum_{k\in s}\, w_k\, x_{yk}}{\sum_{k\in s}\, w_k\, x_{zk}} = \frac{\sum_{k=1}^{N}\, x_{yk}}{\sum_{k=1}^{N}\, x_{zk}} = \frac{t_{xy}}{t_{xz}} = R_0. \tag{9}$$

Proposition 1 (Plikusas, 2001). *The weights w_k of the calibrated estimator of a ratio, which minimize* (8) *with $q_k \equiv 1$ and satisfy* (9)*, can be expressed as*

$$w_k = d_k\left(1 - \frac{\sum_{l\in s} d_l(x_{yl} - R_0 x_{zl})}{\sum_{l\in s} d_l(x_{yl} - R_0 x_{zl})^2}(x_{yk} - R_0 x_{zk})\right), \quad k \in s.$$

The calibrated estimator $\widehat{R}^{(cal)}$ is not unbiased as well as the other estimators of the ratio considered in this paper. It may be written in the form

$$\widehat{R}^{(cal)} = \frac{\widehat{t}_y\widehat{t}_1 - \widehat{t}_2\widehat{t}_3}{\widehat{t}_z\widehat{t}_1 - \widehat{t}_2\widehat{t}_4},$$

with

$$\widehat{t}_1 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})^2, \quad \widehat{t}_2 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk}),$$

$$\widehat{t}_3 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})y_k, \quad \widehat{t}_4 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})z_k.$$

The approximate variance of $\widehat{R}^{(cal)}$ is

$$AVar\,(\widehat{R}^{(cal)}) = \frac{1}{t_z^2} Var\left((\widehat{t}_y - R\widehat{t}_z) + R_{1\,cal}(\widehat{t}_{xy} - R_0\widehat{t}_{xz})\right),$$

here

$$R_{1\,cal} = \frac{Rt_4 - t_3}{t_1}, \quad t_1 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})^2, \quad t_2 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk}) = 0,$$

$$t_3 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})y_k, \quad t_4 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})z_k.$$

The coefficient $R_{1\,cal}$ can be expressed

$$R_{1\,cal} = -\frac{\sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})(y_k - Rz_k)}{\sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})^2}. \tag{10}$$

The expression (3) with

$$v_k = y_k - \widehat{R} z_k + \widehat{R}_{1\,cal}(x_{yk} - R_0 x_{zk}),$$

and

$$\widehat{R}_{1\,cal} = -\frac{\sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})(y_k - \widehat{R} z_k)}{\sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})^2},$$

is used as the estimator of variance of $\widehat{R}^{(cal)}$.

2.5. *Bias of the Calibrated Estimator of a Ratio*

As far as the accuracy of the estimator is measured by the mean square error of the estimator which depends on the variance and bias of the estimator, the expression of the bias of the calibrated estimator of a ratio may be of interest.

Let us start from the approximate expression of the bias of simple estimator of the ratio $\widehat{R} = \widehat{t}_y/\widehat{t}_z$. Expansion of this estimator in Taylor series up to the second-order terms gives us the following expression

$$\widehat{R} \approx R + \frac{1}{t_z}(\widehat{t}_y - R\widehat{t}_z) - \frac{1}{t_z^2}(\widehat{t}_y - t_y)(\widehat{t}_z - t_z) + \frac{R}{t_z^2}(\widehat{t}_z - t_z)^2.$$

Denote

$$\widehat{R}_2^{(sim)} = -\frac{1}{t_z^2}(\widehat{t}_y - t_y)(\widehat{t}_z - t_z) + \frac{R}{t_z^2}(\widehat{t}_z - t_z)^2.$$

Then the approximate bias of $\widehat{R}$ can be expressed as

$$Bias\,(\widehat{R}) = \mathbf{E}\widehat{R} - R \approx ABias\,(\widehat{R}) = \mathbf{E}\widehat{R}_2^{(sim)} = -\frac{1}{t_z^2}Cov(\widehat{t}_y, \widehat{t}_z) + \frac{R}{t_z^2}Var\,(\widehat{t}_z).$$

Using the expansion of $\widehat{R}^{(cal)}$ at the point $(\widehat{t}_y, \widehat{t}_z, \widehat{t}_1, \widehat{t}_2, \widehat{t}_3, \widehat{t}_4) = (t_y, t_z, t_1, t_2, t_3, t_4)$, up to the second-order terms, we get the approximation of $\widehat{R}^{(cal)}$:

$$\widehat{R}^{(cal)} \approx R + \widehat{R}_1^{(cal)} + \widehat{R}_2^{(cal)}.$$

Here

$$\widehat{R}_1^{(cal)} = \frac{1}{t_z}\big((\widehat{t}_y - t_y) + R_{1\,cal}(\widehat{t}_2 - t_2) - R(\widehat{t}_z - t_z)\big),$$

$$\begin{aligned}\widehat{R}_2^{(cal)} &= \Big(-\frac{1}{t_z^2}(\widehat{t}_y - t_y)(\widehat{t}_z - t_z) + \frac{R}{t_z^2}(\widehat{t}_z - t_z)^2\Big) + \frac{t_4 R_{1\,cal}}{t_z^2 t_1}(\widehat{t}_2 - t_2)^2 \\ &\quad + \frac{1}{t_z t_1}\Big(\frac{t_4}{t_z}(\widehat{t}_y - t_y) - R_{1\,cal}(\widehat{t}_1 - t_1) - (\widehat{t}_3 - t_3) \\ &\quad - \frac{2Rt_4 - t_3}{t_z}(\widehat{t}_z - t_z) + R(\widehat{t}_4 - t_4)\Big)(\widehat{t}_2 - t_2) \\ &= \widehat{R}_2^{(sim)} + \frac{t_4 R_{1\,cal}}{t_z^2 t_1}(\widehat{t}_2 - t_2)^2 + \frac{1}{t_z t_1}\Big(\frac{t_4}{t_z}\big(\widehat{t}_y - R\widehat{t}_z - (t_y - Rt_z)\big) \\ &\quad - \frac{t_1 R_{1\,cal}}{t_z}(\widehat{t}_z - t_z) - R_{1\,cal}(\widehat{t}_1 - t_1) \\ &\quad - \big(\widehat{t}_3 - R\widehat{t}_4 - (t_3 - Rt_4)\big)\Big)(\widehat{t}_2 - t_2).\end{aligned}$$

Thus, we can find the approximate bias of $\widehat{R}^{(cal)}$

$$\begin{aligned}Bias\,(\widehat{R}^{(cal)}) &= \mathbf{E}\,(\widehat{R}^{(cal)} - R) \approx ABias\,\widehat{R}^{(cal)} = \mathbf{E}\widehat{R}_2^{(cal)} \\ &= \mathbf{E}\,\widehat{R}_2^{(sim)} + \frac{t_4 R_{1\,cal}}{t_z^2 t_1}\mathbf{E}(\widehat{t}_2 - t_2)^2 \\ &\quad + \frac{1}{t_z t_1}\Big(\frac{t_4}{t_z}\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z - (t_y - Rt_z))(\widehat{t}_2 - t_2) \\ &\quad - \frac{t_1 R_{1\,cal}}{t_z}\,\mathbf{E}\,(\widehat{t}_z - t_z)(\widehat{t}_2 - t_2) - R_{1\,cal}\,\mathbf{E}\,(\widehat{t}_1 - t_1)(\widehat{t}_2 - t_2) \\ &\quad - \mathbf{E}\,((\widehat{t}_3 - R\widehat{t}_4) - (t_3 - Rt_4))(\widehat{t}_2 - t_2)\Big).\end{aligned}$$

Hence follows

Proposition 2. *An approximate bias of the calibrated estimator of a ratio can be expressed as follows*

$$\begin{aligned}ABias\,(\widehat{R}^{(cal)}) &= ABias\,(\widehat{R}) + \frac{t_4 R_{1\,cal}}{t_z^2 t_1}Var\,(\widehat{t}_2) + \frac{1}{t_1 t_z}\Big(Cov(\widehat{t}_y - R\widehat{t}_z, \widehat{t}_2) \\ &\quad - R_{1\,cal}\Big(\frac{t_1}{t_z}Cov(\widehat{t}_z, \widehat{t}_2) + Cov(\widehat{t}_1, \widehat{t}_2)\Big) - Cov(\widehat{t}_3 - R\widehat{t}_4, \widehat{t}_2)\Big).\end{aligned}$$

3. Some Properties of the Estimators of a Ratio

The sampling design when any collection of n different elements from the finite population of size N has equal probabilities $1/C_N^n$ to be selected, is called simple random sampling. It is the sampling with equal selection probabilities without replacement. The estimator of the population total t_y

$$\widehat{t}_y = \frac{N}{n}\sum_{k=1}^{n} y_k$$

is unbiased for this sampling design, its variance equals

$$Var\,(\widehat{t}_y) = N^2\Big(1-\frac{n}{N}\Big)\frac{s_y^2}{n},$$

and the covariance of two estimators of totals

$$Cov\,(\widehat{t}_y,\widehat{t}_z) = \mathbf{E}\,(\widehat{t}_y - \mathbf{E}\,\widehat{t}_y)(\widehat{t}_z - \mathbf{E}\,\widehat{t}_z) = N^2\Big(1-\frac{n}{N}\Big)\frac{s_{yz}^2}{n},$$

$$s_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k-\mu_y)^2, \quad s_{yz}^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k-\mu_y)(z_k-\mu_z).$$

1. Comparison of approximate variances of calibrated and simple estimators of a ratio. *The approximate variance of the calibrated estimator of a ratio is not higher than that of the simple estimator of ratio for simple random sampling*:

$$AVar\,(\widehat{R}^{(cal)}) \leqslant AVar\,(\widehat{R})\,.$$

Proof. The difference

$$\begin{aligned} &AVar\,(\widehat{R}^{(cal)}) - AVar\,(\widehat{R}) \\ &\quad = \frac{1}{t_z^2}R_{1cal}\Big(R_{1cal}Var(\widehat{t}_{xy}-R_0\widehat{t}_{xz}) + 2\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy}-R_0\widehat{t}_{xz})\Big). \end{aligned}$$

In the case of simple random sampling from (10) we have

$$R_{1\ cal} = -\frac{s_{x_y-R_0x_z,y-Rz}}{s^2_{x_y-R_0x_z}} = -\frac{\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy}-R_0\widehat{t}_{xz})}{Var\,(\widehat{t}_{xy}-R_0\widehat{t}_{xz})} \tag{11}$$

and

$$AVar\,(\widehat{R}^{(cal)}) - AVar\,(\widehat{R}) = -\frac{R_{1\,cal}^2}{t_z^2}Var\,(\widehat{t}_{xy}-R_0\widehat{t}_{xz}) \leqslant 0.$$

It means that the approximate variance of the calibrated estimator of a ratio is not greater than that of the simple estimator of a ratio.

2. Comparison of the approximate variance of ratio and simple estimators of a ratio. *Under the condition*

$$\rho\big(\widehat{t}_y - R\widehat{t}_z, \widehat{t}_{xy}-R_0\widehat{t}_{xz}\big) \geqslant \frac{1}{2}\frac{t_y}{t_{xy}}\sqrt{\frac{Var(\widehat{t}_{xy}-R_0\widehat{t}_{xz})}{Var(\widehat{t}_y - R\widehat{t}_z)}} \tag{12}$$

the approximate variance of the ratio estimator of a ratio is not higher than the variance of the simple estimator of a ratio

$$AVar\,(\widehat{R}^{(rat)}) \leqslant AVar\,(\widehat{R})$$

for any sampling design.

Proof. The difference of approximate variances can be expressed as follows:

$$AVar\,(\widehat{R}^{(rat)}) - AVar\,(\widehat{R}) = \frac{1}{t_z^2}\frac{t_y}{t_{xy}} Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \times \left(\frac{t_y}{t_{xy}} - 2\sqrt{\frac{Var\,(\widehat{t}_y - R\widehat{t}_z)}{Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz})}}\,\rho\big(\widehat{t}_y - R\widehat{t}_z, \widehat{t}_{xy} - R_0\widehat{t}_{xz}\big)\right). \quad (13)$$

Solving the inequality $AVar\,(\widehat{R}^{(rat)}) - AVar\,(\widehat{R}) \leqslant 0$ with respect to the correlation coefficient ρ we get (12).

Conclusion. For the simple random sampling we have

$$AVar\,(\widehat{R}^{(rat)}) - AVar\,(\widehat{R}) \leqslant 0$$

if

$$\frac{t_y}{t_{xy}} Var\,(\widehat{t}_{xy} - R_0\widehat{t}xz) - 2\mathbf{E}(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \leqslant 0$$

because of (13). This condition can be expressed as

$$N^2\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\Big(\frac{t_y}{t_{xy}} s^2_{x_y - R_0 x_z} - 2s_{y-Rz, x_y - R_0 x_z}\Big) \leqslant 0,$$

or

$$A(y, z, x_y, x_z) = \frac{t_y}{t_{xy}}\sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})^2 - 2\sum_{k=1}^{N}(y_k - Rz_k)(x_{yk} - R_0 x_{zk}) \leqslant 0. \quad (14)$$

3. Comparison of approximate variances of calibrated and ratio estimators of a ratio. *The approximate variance of the calibrated estimator of a ratio is not higher than the approximate variance of the ratio estimator of a ratio for simple random sampling*:

$$AVar\,(\widehat{R}^{(cal)}) \leqslant AVar\,(\widehat{R}^{(rat)}). \quad (15)$$

Proof. The difference of approximate variances can be expressed as

$$\begin{aligned} &AVar\,(\widehat{R}^{(cal)}) - AVar\,(\widehat{R}^{(rat)}) \\ &\quad = \frac{1}{t_z^2}\Big(R^2_{1\,cal} Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) + 2R_{1\,cal}\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \\ &\qquad + 2R_{1\,rat}\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) - R^2_{1\,rat} Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz})\Big) \\ &\quad = \frac{R_{1\,cal} + R_{1\,rat}}{t_z^2} Var(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \end{aligned}$$

$$\times \left((R_{1\,cal} - R_{1\,rat}) + 2\frac{\mathbf{E}\,(\widehat{t}_y - R\widehat{t}_z)(\widehat{t}_{xy} - R_0\widehat{t}_{xz})}{Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz})} \right)$$

$$= \frac{R_{1\,cal} + R_{1\,rat}}{t_z^2} Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \Bigg(R_{1\,cal} - R_{1\,rat}$$

$$+ 2\sqrt{\frac{Var\,(\widehat{t}_y - R\widehat{t}_z)}{Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz})}}\,\rho\,(\widehat{t}_y - R\widehat{t}_z, \widehat{t}_{xy} - R_0\widehat{t}_{xz}) \Bigg).$$

Thus, we have $AVar\,(\widehat{R}^{(cal)}) \leqslant AVar\,(\widehat{R}^{(rat)})$, if

$$\rho(\widehat{t}_y - R\widehat{t}_z, \widehat{t}_{xy} - R_0\widehat{t}_{xz}) \leqslant \frac{1}{2}(R_{1\,rat} - R_{1\,cal})\sqrt{\frac{Var\,(\widehat{t}_{xy} - R_0\widehat{t}_{xz})}{Var\,(\widehat{t}_y - R\widehat{t}_z)}}$$

for any sampling design.

In the case of simple random sampling, we obtain

$$AVar\,(\widehat{R}^{(cal)}) - AVar\,(\widehat{R}^{(rat)})$$

$$= \frac{R_{1\,cal} + R_{1\,rat}}{t_z^2} Var(\widehat{t}_{xy} - R_0\widehat{t}_{xz})(R_{1\,cal} - R_{1\,rat} - 2R_{1\,cal})$$

$$= -\frac{(R_{1\,cal} + R_{1\,rat})^2}{t_z^2} Var(\widehat{t}_{xy} - R_0\widehat{t}_{xz}) \leqslant 0,$$

it means that the variance of the calibrated estimator of the ratio is not higher than the variance of the ratio estimator of a ratio.

4. Simulation Study

The data of the artificial populations of size $N = 100$ were used for the simulation study. Three collections of variables were generated:

Case 1. Non-correlated variables:

$$\rho(y, z) = 0.09, \quad \rho(y, x_y) = 0.11, \quad \rho(z, x_z) = 0.1, \quad \rho(x_y, x_z) = 0.1.$$

An unknown ratio of the totals of study variables $R = t_y/t_z = 3.3477$ which has to be estimated, the known ratio of the totals of auxiliary variables $R_0 = t_{xy}/t_{xz} = 0.9711$.

Case 2. Non-correlated study variables, and study variables correlated with the auxiliary variables:

$$\rho(y, z) = 0.09, \quad \rho(y, x_y) = 0.83, \quad \rho(z, x_z) = 0.83, \quad \rho(x_y, x_z) = 0.04.$$

Parameters $R = 3.3477$, $R_0 = 3,7378$.

Case 3. Correlated study variables, and study variables correlated with the auxiliary variables:

$$\rho(y, z) = 0.8, \quad \rho(y, x_y) = 0.8, \quad \rho(z, x_z) = 0,8, \quad \rho(x_y, x_z) = 0.59.$$

Parameters $R = 3.8825$, $R_0 = 3.9957$.

The values of variables y and z of the artificial population are presented in Fig. 1 and Fig. 2.

1000 simple random samples of sizes n =10, 20, 30 were drawn from the population. The estimates $\widehat{R}$, $\widehat{R}^{(rat)}$, $\widehat{R}^{(reg)}$, $\widehat{R}^{(cal)}$ as well as estimates of their approximate variances have been calculated in each case.

The results of simulation are presented in Tables 1, 2, 3. For all estimators $\widehat{\theta} = \widehat{R}$, $\widehat{R}^{(rat)}$, $\widehat{R}^{(reg)}$, $\widehat{R}^{(cal)}$: averages of the estimates

$$\bar{\widehat{\theta}} = \frac{1}{1000} \sum \widehat{\theta},$$

estimated biases $\widehat{Bias}\,(\widehat{\theta}) = \bar{\widehat{\theta}} - R$, approximate variances, estimated mean square errors $\widehat{MSE}\,(\widehat{\theta}) = AVar\,(\widehat{\theta}) + (\widehat{Bias}\,(\widehat{\theta}))^2$, estimated relative mean square errors

$$Rel\,\widehat{MSE}\,(\widehat{\theta}) = \frac{\sqrt{\widehat{MSE}\,(\widehat{\theta})}}{\bar{\widehat{\theta}}} 100(\%),$$

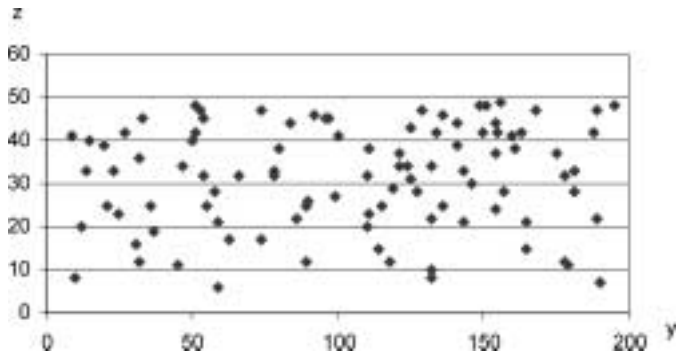

Fig. 1. Scatter-plot of the variables y and z in Cases 1, 2.

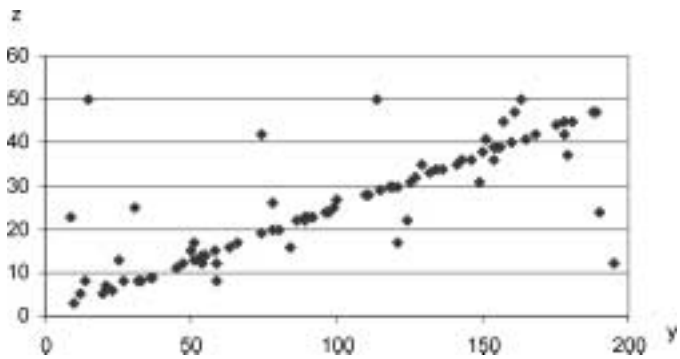

Fig. 2. Scatter-plot of the variables y and z in Case 3.

Table 1

Results of the estimation in Case 1

Estimator $\widehat{\theta}$	Sample size n	Average estimate	Estimated bias	Approximate variance	$\widehat{MSE}$	$\frac{Rel}{\widehat{MSE}}$ (%)	Average estimate of variance
$\widehat{R}$	10	3.3784	0.0307	0.3667	0.3676	17.95	0.4089
	20	3.3439	−0.0038	0.1630	0.1630	12.07	0.1704
	30	3.3465	−0.0012	0.0951	0.0951	9.22	0.0974
$\widehat{R}^{(rat)}$	10	3.3487	0.0890	0.6732	0.6811	24.01	0.7701
	20	3.3775	0.0298	0.2992	0.3001	16.22	0.3179
	30	3.3791	0.0314	0.1745	0.1755	12.40	0.1820
$\widehat{R}^{(reg)}$	10	3.4160	0.0683	0.3556	0.3603	17.57	0.3542
	20	3.3468	−0.0009	0.1581	0.1581	11.88	0.1579
	30	3.3521	0.0044	0.0922	0.0922	9.06	0.0926
$\widehat{R}^{(cal)}$	10	3.3702	0.0255	0.3520	0.3525	17.62	0.3462
	20	3.3358	−0.0119	0.1565	0.1566	11.86	0.1557
	30	3.3454	−0.0023	0.0913	0.0329	9.03	0.0917

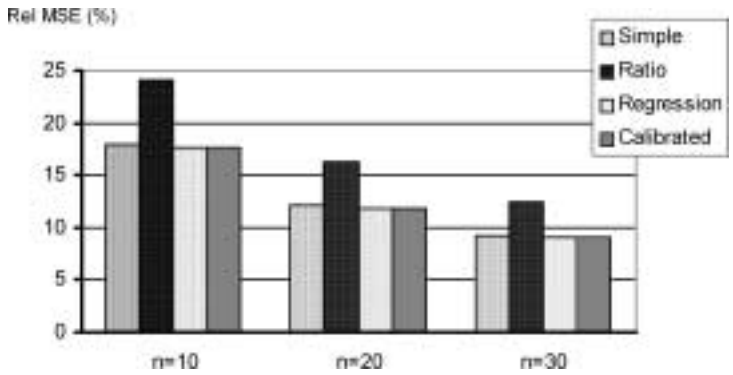

Fig. 3. Results of the estimation of the ratio in Case 1.

averages of the estimates of variances

$$\overline{\widehat{Var}}\,(\widehat{\theta}) = \frac{1}{1000}\sum \widehat{Var}\,(\widehat{\theta})$$

are shown in the tables. The dependence of the estimated mean square errors of the estimates on the sample size n is presented in Figs. 3, 4, 5.

Table 2
Results of the estimation in Case 2

Estimator $\widehat{\theta}$	Sample size n	Average estimate	Estimated bias	Approximate variance	$\widehat{MSE}$	$\frac{Rel}{\widehat{MSE}}$ (%)	Average estimate of variance
$\widehat{R}$	10	3.3783	0.0306	0.3667	0.3676	17.95	0.4089
	20	3.3439	−0.0038	0.1630	0.1630	12.07	0.1704
	30	3.3465	−0.0012	0.0951	0.0951	9.22	0.0974
$\widehat{R}^{(rat)}$	10	3.3530	0.0053	0.1192	0.1192	10.30	0.1332
	20	3.3625	0.0148	0.0530	0.0532	6.86	0.0535
	30	3.3486	0.0009	0.0309	0.0309	5.25	0.0312
$\widehat{R}^{(reg)}$	10	3.3313	−0.0164	0.1037	0.1040	9.68	0.1038
	20	3.3504	0.0027	0.0461	0.0461	6.41	0.0446
	30	3.3420	−0.0057	0.0269	0.0269	4.91	0.0264
$\widehat{R}^{(cal)}$	10	3.3468	−0.0009	0.1027	0.1027	9.58	0.1026
	20	3.3550	0.0073	0.0457	0.0458	6.38	0.0441
	30	3.3484	0.0007	0.0266	0.0266	4.87	0.0261

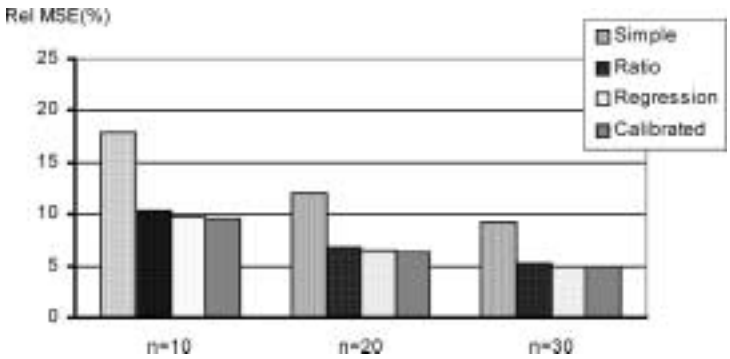

Fig. 4. Results of the estimation of the ratio in Case 2.

One can see comparatively high mean square errors of the ratio estimator of a ratio in Cases 1 and 3. Only in Case 2 it becomes less than that of the simple estimator of a ratio. This phenomena can be explained by checking the condition (14) (see Table 4).

Condition (14) is only satisfied in Case 2, and only in this case the approximate variance of the ratio estimator of ratio $R^{(rat)}$ is less than the approximate variance of the simple estimator of the ratio $\widehat{R}$.

Table 3

Results of the estimation in Case 3

Estimator $\widehat{\theta}$	Sample size n	Average estimate	Estimated bias	Approximate variance	$\widehat{MSE}$	$\frac{Rel}{\widehat{MSE}}$ (%)	Average estimate of variance
$\widehat{R}$	10	3.8948	0.0123	0.1270	0.1272	9.16	0.1297
	20	3.8937	0.0112	0.0564	0.0565	6.11	0.0582
	30	3.8783	−0.0042	0.0329	0.0329	4.68	0.0337
$\widehat{R}^{(rat)}$	10	3.9125	0.0300	0.1448	0.1457	9.76	0.1731
	20	3.9144	0.0319	0.0644	0.0654	6.53	0.0704
	30	3.8898	−0.0073	0.-375	0.0376	4.98	0.0394
$\widehat{R}^{(reg)}$	10	3.8586	−0.0239	0.1057	0.1063	8.45	0.1122
	20	3.8877	0.0052	0.0470	0.0470	5.58	0.0486
	30	3.8767	−0.0058	0.0274	0.0274	4.27	0.0281
$\widehat{R}^{(cal)}$	10	3.8841	0.0016	0.0733	0.0733	6.97	0.0473
	20	3.827	0.0102	0.0326	0.0327	4.65	0.0269
	30	3.8801	−0.0024	0.0190	0.0190	3.55	0.0169

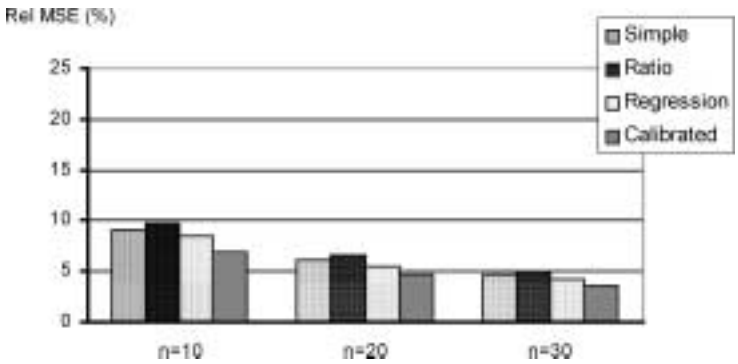

Fig. 5. Results of the estimation of the ratio in Case 3.

5. Conclusions

1. The calibrated estimator of a ratio has the smallest mean square error among the estimators investigated.
2. The simulation results do not show any significant bias of the estimates.
3. The ratio estimator of a ratio has to be applied carefully because its mean square error can be larger than that of the simple estimator. Even in Case 3, when all variables are correlated, the condition (12) is not satisfied.
4. The method of calibration of weights of the estimator of a ratio does not guarantee the positivity of weights w_k. The case of the negative weights is not attractive for practitioners because it is thought that a sampled element could represent at least

Table 4
Statistic for comparison of the simple and the ratio estimates of a ratio

Case	1	2	3
$A(y, z, x_y, x_z)$	330 030	−287 598	15 261

itself and its weight has to be greater or equal to 1. So, restrictions on the weights can be put, but then the properties of the estimators would change.

5. The problem of calibration of weights may be generalized for the estimation of some other nonlinear functions of totals. Some other distance measures between the design weights and the calibrated weights can be used.

Acknowledgements. The authors are thankful to Dalius Pumputis for the data of artificial population, to the Swedish Institute and Umeå University for support, and to Prof. Gunnar Kulldorff for his co-operation when working at this paper.

References

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, New York.

Deville, J.-C., and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

Plikusas, A. (2001). Calibrated estimators of the ratio. *Lithuanian Math. J.*, **41**, 457–462.

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

D. Krapavickaitė received her PhD degree from the Vilnius University in 1981. She is a senior researcher of the Institute of Mathematics and Informatics and a docent of Vilnius Gediminas Technical University. Her recent research interests include survey sampling and statistical analysis of data.

A. Plikusas received his PhD degree from the Vilnius University in 1980. He is a senior researcher of the Institute of Mathematics and Informatics and a docent of Vilnius University. His recent research interests include finite population statistics.

Santykio vertinimas baigtinėje populiacijoje

Danutė KRAPAVICKAITĖ, Aleksandras PLIKUSAS

Nagrinėjami dviejų sumų santykio vertinimo būdai, turint iš baigtinės populiacijos išrinktą tikimybinę imtį. Lyginami 4 santykio įvertiniai: paprastasis – Horvico–Tompsono sumų įvertinių santykis, santykinis – santykinių sumos įvertinių santykis, regresinis – regresinių sumos įvertinių santykis ir kalibruotas santykio įvertinys. Tiriamos šių santykio įvertinių savybės. Teoriniai rezultatai iliustruojami modeliavimo pavyzdžiu.