Nonlinear Analysis: Modelling and Control, 2005, Vol. 10, No. 4, 333–342

Investigation of Accuracy of a Calibrated Estimator of a Ratio by Modelling

V. Chadyšas[1], **D. Krapavickaitė**[1,2]

[1]Vilnius Gediminas Technical University
4040@one.lt
[2]Institute of Mathematics and Informatics
krapav@ktl.mii.lt

Received: 10.10.2005
Accepted: 24.10.2005

Abstract. Estimator of finite population parameter – ratio of totals of two variables – is investigated by modelling in the case of simple random sampling. Traditional estimator of the ratio is compared with the calibrated estimator of the ratio introduced by Plikusas [1]. The Taylor series expansion of the estimators are used for the expressions of approximate biases and approximate variances [2]. Some estimator of bias is introduced in this paper. Using data of artificial population the accuracy of two estimators of the ratio is compared by modelling. Dependence of the estimates of mean square error of the estimators of the ratio on the correlation coefficient of variables which are used in the numerator and denominator, is also shown in the modelling.

Keywords: finite population, ratio of two totals, simple random sampling, calibrated estimator.

1 Introduction

The ratio of totals of two study variables in finite population is investigated. This parameter is often met in official statistics. Calibration of the total is commonly used in order to get higher accuracy of the estimators using auxiliary information.

The idea of calibration of the estimators of totals was presented by Deville and Särndal [3]. A calibrated estimator of the ratio of two totals was introduced and its approximate variance given by Plikusas [1]. The aim of this paper is to introduce an estimator of bias of the calibrated estimator of the ratio, and to show

the accuracy of this estimator depending on the correlation coefficient between the variables which are used in the numerator and denominator of the ratio, by modelling.

2 Estimators of the ratio

2.1 Traditional estimator of the ratio

Suppose $\mathcal{U} = \{1, \dots, N\}$ is a finite population, y and z are two study variables defined for the elements of the population $\mathcal{U}$ with the unknown values $\{y_1, \dots, y_N\}$ and $\{z_1, \dots, z_N\}$, respectively. 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

$$R = \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{N} z_k} = \frac{t_y}{t_z} \tag{1}$$

in the case of simple random sampling.

Simple random sampling (SRS) is a sampling design in which all possible collections of n different elements s, $s \subset \mathcal{U}$, have the same probability $1/C_N^n$ of selection ([4]). The SRS design may be obtained when n elements from the finite population are drawn with equal selection probabilities without replacement. In the case of SRS, the estimator of population total t_y of any variable y

$$\hat{t}_y = \frac{N}{n} \sum_{k \in s} y_k = \sum_{k \in s} d_k y_k$$

is unbiased. $d_k = N/n$, $k \in s$ are called the SRS design weights. The variance of the estimator $\hat{t}_y$ is

$$Var(\hat{t}_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{s_y^2}{n}, \quad s_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} \Big(y_k - \frac{t_y}{N}\Big)^2,$$

an estimator of variance

$$\widehat{Var}(\hat{t}_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{\hat{s}_y^2}{n}, \quad \hat{s}_y^2 = \frac{1}{n-1} \sum_{k \in s} \Big(y_k - \frac{\hat{t}_y}{N}\Big)^2$$

is unbiased. Covariance of two estimators of totals is

$$Cov(\hat{t}_y, \hat{t}_z) = N^2\Big(1 - \frac{n}{N}\Big)\frac{s^2_{yz}}{n},$$
$$s^2_{yz} = \frac{1}{N-1}\sum_{k=1}^{N}\Big(y_k - \frac{t_y}{N}\Big)\Big(z_k - \frac{t_z}{N}\Big),$$

its estimator

$$\widehat{Cov}(\hat{t}_y, \hat{t}_z) = N^2\Big(1 - \frac{n}{N}\Big)\frac{\hat{s}^2_{yz}}{n},$$
$$\hat{s}^2_{yz} = \frac{1}{n-1}\sum_{k\in s}\Big(y_k - \frac{\hat{t}_y}{N}\Big)\Big(z_k - \frac{\hat{t}_z}{N}\Big)$$

is also unbiased. Referring to [2], we have the following proposition.

Proposition 1. *In the case of SRS the bias* $Bias(\widehat{R}) = \boldsymbol{E}\widehat{R} - R$ *of the estimator*

$$\widehat{R} = \hat{t}_y/\hat{t}_z \tag{2}$$

of the ratio (1) *is expressed approximately as*

$$ABias(\widehat{R}) = \frac{1}{t_z^2}\big(R\,Var(\hat{t}_z) - \mathrm{Cov}(\hat{t}_y, \hat{t}_z)\big).$$

The approximate variance of (2) *equals*

$$AVar(\widehat{R}) = \frac{1}{t_z^2}N^2\Big(1 - \frac{n}{N}\Big)\frac{s_d^2}{n}, \quad s_d^2 = \frac{1}{N-1}\sum_{k=1}^{N} v_k^2$$

with $v_k = y_k - Rz_k$.

We estimate the variance $Var(\widehat{R})$ by

$$\widehat{Var}(\widehat{R}) = \frac{1}{\hat{t}_z^2}N^2\Big(1 - \frac{n}{N}\Big)\frac{\hat{s}_d^2}{n}, \quad \hat{s}_d^2 = \frac{1}{n-1}\sum_{k\in s}\hat{v}_k^2 \tag{3}$$

with $\hat{v}_k = y_k - \widehat{R}z_k$, and the bias $Bias(\widehat{R})$ by

$$\widehat{Bias}(\widehat{R}) = \frac{1}{\hat{t}_z^2}\big(\widehat{R}\,\widehat{Var}(\hat{t}_z) - \widehat{Cov}(\hat{t}_y, \hat{t}_z)\big).$$

2.2 Calibrated estimator of the ratio

Suppose a variable x_y with the population values $x_{y1}, \ldots, x_{yN}$ and a variable x_z with the values $x_{z1}, \ldots, x_{zN}$ are auxiliary variables with totals

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

The estimator of the ratio R

$$\widehat{R}^{(cal)} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k z_k}$$

is said to be calibrated (Plikusas, [1]) if the new weights w_k minimize the function

$$L(w,d) = \sum_{k\in s} \frac{(w_k - d_k)^2}{d_k}$$

and estimate the known ratio $R_0 = t_{xy}/t_{xz}$ of totals of the auxiliary variables x_y and x_z without error:

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

The approximate variance $AVar\widehat{R}^{(cal)}$ and the approximate bias $ABias(\widehat{R}^{(cal)})$ obtained using Taylor series expansion of the estimator $\widehat{R}^{(cal)}$, are presented in [2] for any sampling design. We restrict ourselves with SRS.

Proposition 2. *The calibrated estimator $\widehat{R}^{(cal)}$ in SRS can be written in the form*

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

with

$$\hat{t}_1 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})^2, \quad \hat{t}_2 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk}),$$
$$\hat{t}_3 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk}) y_k, \quad \hat{t}_4 = \sum_{k\in s} d_k(x_{yk} - R_0 x_{zk}) z_k.$$

The approximate variance of $\widehat{R}^{(cal)}$ can be written

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

here $R_{1\,cal} = (Rt_4 - t_3)/t_1$,

$$t_1 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})^2, \quad t_2 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk}) = 0,$$
$$t_3 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})y_k, \quad t_4 = \sum_{k=1}^{N}(x_{yk} - R_0 x_{zk})z_k.$$

The approximate bias of $\widehat{R}^{(cal)}$ can be expressed as follows

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

We will use the expression (3) with

$$\hat{v}_k = y_k - \widehat{R}z_k + \widehat{R}_{1\,cal}(x_{yk} - R_0 x_{zk}),$$
$$\widehat{R}_{1\,cal} = (\widehat{R}\hat{t}_4 - \hat{t}_3)/\hat{t}_1,$$

as an estimator $\widehat{Var}(\widehat{R}^{(cal)})$ of $Var(\widehat{R}^{(cal)})$.

For the estimation of bias $Bias(\widehat{R}^{(cal)})$ we introduce the estimator

$$\widehat{Bias}(\widehat{R}^{(cal)}) = \frac{1}{\hat{t}_z^2}\big(\widehat{R}^{(cal)}\widehat{Var}(\hat{t}_z) - \widehat{Cov}(\hat{t}_y, \hat{t}_z)\big) + \frac{\hat{t}_4\widehat{R}_{1\,cal}}{\hat{t}_z^2\hat{t}_1}\widehat{Var}(\hat{t}_2)$$
$$+ \frac{1}{\hat{t}_1\hat{t}_2}\bigg(N^2\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\frac{1}{n-1}\sum_{k\in s}(y_k - \widehat{R}_{1\,cal}z_k)(x_{yk} - R_0 x_{zk})$$
$$- \widehat{R}_{1\,cal}\Big(\frac{\hat{t}_1}{\hat{t}_z}\widehat{Cov}(\hat{t}_z, \hat{t}_2) + \widehat{Cov}(\hat{t}_1, \hat{t}_2)\Big) - N^2\Big(1 - \frac{n}{N}\Big)\frac{1}{n}\frac{1}{n-1}$$
$$\cdot \sum_{k\in s}\big((x_{yk} - R_0 x_{zk})y_k - \widehat{R}^{(cal)}(x_{yk} - R_0 x_{zk})z_k\big)(x_{yk} - R_0 x_{zk})\bigg).$$

3 Results of modelling

The data of artificial population of size $N = 87$ was used for the simulation study. Two collections of variables were generated.

Case 1. Highly correlated study variables y, z and highly correlated auxiliary variables x_y, x_z with the correlation coefficients

$$\rho(y,z) = \frac{1}{N-1}\frac{\sum_{k=1}^{N}(y_k - t_y/N)(z_k - t_z/N)}{s_y s_z} = 0.98,$$

$\rho(y,x_y) = 0.94$, $\rho(z,x_z) = 0.95$, $\rho(x_y,x_z) = 0.90$ and the population variances $\sigma_y^2 = 145\,270\,631$, $\sigma_z^2 = 149$, $\sigma_{x_y}^2 = 4.8 \cdot 10^{10}$, $\sigma_{x_z}^2 = 173$. The notation $\sigma_u^2 = (N-1)s_u^2/N$ is used here for $u = y,\ z,\ x_y,\ x_z$.

Case 2. Low-correlated study variables with the correlation coefficients $\rho(y,z) = 0.27$, $\rho(y,x_y) = 0.94$, $\rho(z,x_z) = 0.95$, $\rho(x_y,x_z) = 0.27$ and having the same population variances as in Case 1.

An unknown ratio of totals of the study variables, $R = 2\,227$, has to be estimated. 1 000 simple random samples of size $n = 10,\ 20,\ 30$ have been drawn from the artificial population. The estimates $\widehat{R}$ and $\widehat{R}^{(cal)}$ as well as their approximate variances and approximate biases have been calculated in each case. The results of simulation are presented in Tables 1, 2. For both estimators $\hat{\theta} = \widehat{R}$ and $\hat{\theta} = \widehat{R}^{(cal)}$ the average of the estimates

$$\bar{\hat{\theta}} = \frac{1}{1\,000}\sum_{k=1}^{1\,000}\hat{\theta}_k$$

(replicates of $\hat{\theta}$ are denoted by $\hat{\theta}_k$), the empirical variance of the estimates

$$\widetilde{Var}(\hat{\theta})^2 = \frac{1}{1\,000}\sum_{k=1}^{1\,000}(\hat{\theta}_k - \bar{\hat{\theta}})^2,$$

the average of the estimates of variances

$$\overline{\widehat{Var}}(\hat{\theta}) = \frac{1}{1\,000}\sum_{k=1}^{1\,000}\widehat{Var}(\hat{\theta}_k),$$

the average of the estimates of the biases

$$\overline{\widehat{Bias}}(\hat{\theta}) = \frac{1}{1\,000}\sum_{k=1}^{1\,000}\widehat{Bias}(\hat{\theta}_k),$$

and the estimated mean square errors

$$\widehat{MSE}(\hat{\theta}) = \overline{\widehat{Var}}(\hat{\theta}) + \left(\overline{\widehat{Bias}}(\hat{\theta})\right)^2$$

are calculated.

The true values of approximate variance and approximate bias as well as the simulated averages of the estimates of variance and bias are given in Tables 1, 2. The dependence of the estimated variances of estimates on the sample size is presented in Fig. 1. The dependence of the estimated biases of the estimates on the sample size is presented in Fig. 2. The dependence of the mean squared errors of the estimates on the sample size is presented in Fig. 3. The dependence of approximate variances, empirical variances and averages of the estimated variances on the sample size is presented in Figs. 4, 5.

Table 1. The results of estimation in Case 1

$\hat{\theta}$	n	$\bar{\hat{\theta}}$	$AVar(\hat{\theta})$	$\overline{\widehat{Var}}(\hat{\theta})$	$ABias(\hat{\theta})$	$\overline{\widehat{Bias}}(\hat{\theta})$	$\widehat{MSE}(\hat{\theta})$
$\widehat{R}$	10	2 266	39 587	46 900	29.8	32.9	47 980
	20	2 237	16 967	18 082	12.8	13.3	18 257
	30	2 236	9 425	9 855	7.1	7.3	9 908
$\widehat{R}^{(cal)}$	10	2 282	33 548	38 350	31.2	43.4	40 237
	20	2 239	14 378	14 978	13.4	16.3	15 243
	30	2 234	7 988	8 247	7.4	8.8	8 321

Table 2. The results of estimation in Case 2

$\hat{\theta}$	n	$\bar{\hat{\theta}}$	$AVar(\hat{\theta})$	$\overline{\widehat{Var}}(\hat{\theta})$	$ABias(\hat{\theta})$	$\overline{\widehat{Bias}}(\hat{\theta})$	$\widehat{MSE}(\hat{\theta})$
$\widehat{R}$	10	2 296	110 668	133 516	45.8	51.7	136 186
	20	2 243	47 429	50 656	19.6	20.4	51 071
	30	2 232	26 350	27 331	10.9	11.2	27 456
$\widehat{R}^{(cal)}$	10	2 291	48 055	52 592	60.5	63.6	56 642
	20	2 245	20 595	21 082	25.9	26.4	21 781
	30	2 235	11 442	11 522	14.4	14.7	11 737

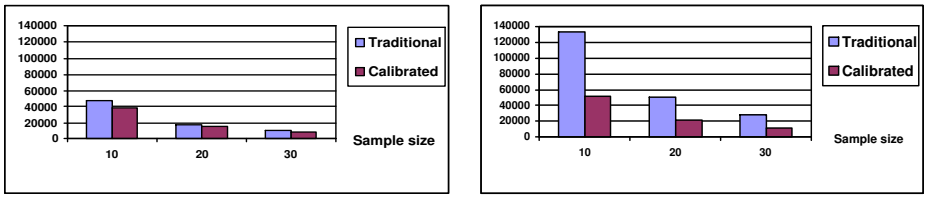

Fig. 1. Estimated variances in Case 1 and Case 2.

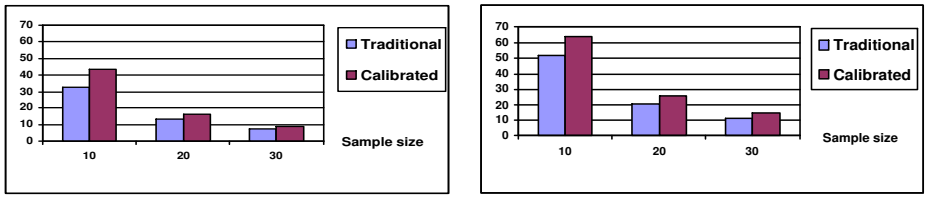

Fig. 2. Estimated biases in Case 1 and Case 2.

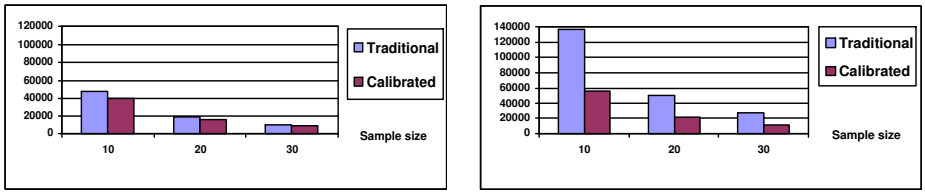

Fig. 3. Estimated mean squared errors in Case 1 and Case 2.

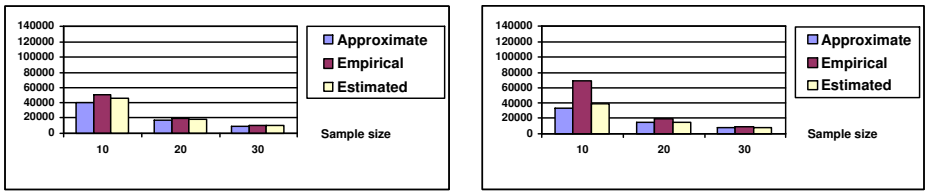

Fig. 4. Variances of the traditional and the calibrated estimator of the ratio in Case 1.

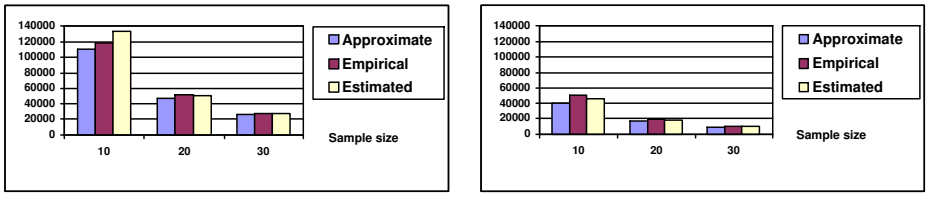

Fig. 5. Variances of the traditional and the calibrated estimator of the ratio in Case 2.

4 Conclusions

The simulation results show that in the case of SRS

1. The approximate bias of the calibrated estimator of the ratio is larger than that of the traditional estimator. The approximate bias of the calibrated estimator of the ratio as well as of the traditional estimator is increasing when the correlation coefficient between the variables used in the numerator and denominator of the ratio is decreasing.

2. The approximate variance of the traditional estimator of the ratio is increasing when the correlation coefficient of the variables in the numerator and the denominator of the ratio is decreasing. The approximate variance of the calibrated estimator of the ratio is not so sensitive to this coefficient of correlation.

3. The estimated mean squared error of the traditional estimator of the ratio is increasing when the correlation coefficient of the variables used in the numerator and denominator of the ratio is decreasing. The estimated mean squared error of the calibrated estimator is not so sensitive to this coefficient of correlation.

4. In simple random sampling the calibrated estimator of the ratio is more efficient than the traditional one when the correlation coefficient between variables used in the numerator and denominator is small.

5. The approximate variance and average estimate of the variance of the calibrated estimator of the ratio are smaller than the empirical variance for small sample size. If higher order terms of Taylor expansion would be taken into expression of the approximate variance of this estimator, one can expect to improve the accuracy of the approximation of the variance.

References

1. A. Plikusas. Calibrated Estimators of the Ratio, *Lietuvos matematikos rinkinys*, **41**, spec. Nr., pp. 457–462, 2001.

2. D. Krapavickaitė, A. Plikusas. Estimation of a Ratio in a Finite Population, *Informatica*, **16**(3), pp. 347–364, 2005.

3. J. Deville, C.-E. Särndal. Calibrated Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, pp. 376–382, 1992.

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