Liet. matem. rink., **44**, *spec. nr.*, 2004, 572–576

Calibrated estimators of totals under different distance measures

Aleksandras PLIKUSAS* (MII), Dalius PUMPUTIS (VPU)
e-mail: plikusas@ktl.mii.lt, dalpas@delfi.lt

1. Introduction

Calibrated estimators are widely used to improve the quality of estimators, using auxiliary information, in finite population statistics. The idea of calibration technique for estimating of population totals was presented in [1]. Recently the calibration technique has been widely used in the presence of nonresponse ([2], [3]). The calibrated estimators of the ratio of two totals were introduced in [6]. Five distance functions have been presented in [1], but only one of them (L_1) is used in practice. This distance function is the simplest one and there exists an explicit solution of calibration equations when calibrating the estimator of the ratio [5]. An undesirable property of this distance function is that for some populations calibrated weights can be negative. In this article the performance of some other distance measures has been studied. Some simulation results are presented.

2. Calibration problem

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements and a population variable y taking values $y_1, \ldots, y_N$. We are interested in the estimation of the total

$$t = \sum_{k=1}^{N} y_k.$$

Let us consider the Horvitz–Thompson estimator of the total

$$\hat{t}_\pi = \sum_{k\in s} \frac{y_k}{\pi_k} = \sum_{k\in s} d_k y_k.$$

Here s denotes a probability sample set, $d_k = 1/\pi_k$ are sample design weights, π_k is a probability of inclusion of the element k into the sample s.

Suppose that, for each population element k, $k = 1, \ldots, N$, the vector of the auxiliary variable $\mathbf{x}'_k = (x_{k1}, \ldots, x_{kj})$ is known and denote the known total as $\mathbf{t}_x = \sum_{k=1}^{N} \mathbf{x}_k$.

*The research is supported by The Lithuanian State Science and Studies Foundation (Grant No. T-04051).

Using this auxiliary variable the calibrated estimator

$$\hat{t}_w = \sum_{k \in s} w_k \, y_k$$

of the total t is defined under the following conditions:

a) the weights w_k estimate the known total $\mathbf{t}_x$ without error:

$$\hat{\mathbf{t}}_x = \sum_{k \in s} w_k \mathbf{x}_k = \mathbf{t}_x,$$

b) the distance between the design weights d_k and calibrated weights w_k is minimal according to some loss function L.

It is known that in case the auxiliary variable is well correlated with study variable y, the variance of the calibrated estimator of the total is lower.

3. Examples of distance measures

Let us introduce free additional weights q_k, $k = 1, \ldots, N$. One can modify calibrated estimators by choosing q_k. A number of known estimators can be derived as a special case of the calibrated estimator by choosing weights q_k. Otherwise we can put $q_k = 1$ for all k. The following loss functions can be considered:

$$L_1 = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k q_k}, \quad L_2 = \sum_{k \in s} \frac{w_k}{q_k} \log \frac{w_k}{d_k} - \frac{1}{q_k}(w_k - d_k),$$

$$L_3 = \sum_{k \in s} 2 \frac{(\sqrt{w_k} - \sqrt{d_k})^2}{q_k}, \quad L_4 = \sum_{k \in s} -\frac{d_k}{q_k} \log \frac{w_k}{d_k} + \frac{1}{q_k}(w_k - d_k),$$

$$L_5 = \sum_{k \in s} \frac{(w_k - d_k)^2}{w_k q_k}, \quad L_6 = \sum_{k \in s} \frac{1}{q_k} \left(\frac{w_k}{d_k} - 1 \right)^2, \quad L_7 = \sum_{k \in s} \frac{1}{q_k} \left(\frac{\sqrt{w_k}}{\sqrt{d_k}} - 1 \right)^2.$$

The functions L_1–L_5 are mentioned in [1]. The distance measures L_6 and L_7 are introduced in [6].

4. Results

It has been proved in [1], that the calibrated weights w_k which satisfy the calibration equation $\mathbf{t}_x = \hat{\mathbf{t}}_x$ and minimize the loss function L_1 can be expressed as $w_k = d_k v_k^{(1)}$, where

$$v_k^{(1)} = 1 + q_k \left(\sum_{k=1}^{N} \mathbf{x}'_k - \sum_{k \in s} d_k \mathbf{x}'_k \right) \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}'_k q_k d_k \right)^{-1} \mathbf{x}_k.$$

We will present the corresponding results for some other loss functions.

Table 1. Auxiliary vector $\mathbf{x} = (1, x)$
True value of total: $t = 10652$
Sample size: $n = 20$

Loss function	Estimate of total	Estimated variance	Bias	MSE	cv	$\max\limits_{1 \leqslant k \leqslant m} \lvert d_k - w_k \rvert$	$\frac{1}{m} \sum\limits_{k=1}^{m} \lvert d_k - w_k \rvert$
Coefficient of correlation 0.8							
L_1	10645	490166	−7.360	490220	0.066	3.9234	1.2248
L_3	10666	790586	14.060	790784	0.083	0.0089	0.0050
L_6	10650	509440	−1.710	509443	0.067	3.9841	1.2514
L_7	10659	761272	7.470	761328	0.082	0.0089	0.0050
Coefficient of correlation 0.6							
L_1	10623	955775	−28.56	956591	0.0920	4.0030	1.2272
L_3	10591	940704	−60.86	944407	0.0916	0.0088	0.0050
L_6	10629	890379	−22,70	890894	0.0888	4.0827	1.2721
L_7	10561	953496	−90.44	961675	0.0925	0.0088	0.0050
Coefficient of correlation 0.4							
L_1	10622	1368698	−29.51	1369570	0.1101	4.0263	1.2220
L_3	10596	1194437	−55.28	1197493	0.1031	0.0089	0.0050
L_6	10578	355111	−73.75	1360551	0.0563	4.0327	1.2413
L_7	10661	1237011	9.55	1237102	0.1043	0.0088	0.0050
Coefficient of correlation 0.2							
L_1	10546	1479104	−105.18	1490166	0.1153	3.9780	1.2425
L_3	10663	1306623	11.48	1306755	0.1072	0.0088	0.0050
L_6	10555	1581360	−96.71	1590713	0.1191	3.9108	1.2280
L_7	10625	1402203	−26.02	1402880	0.1114	0.0089	0.0050

PROPOSITION 1. *The weights w_k, which satisfy the calibration equation $\mathbf{t}_x = \hat{\mathbf{t}}_x$ and minimize the loss function L_3 can be expressed as $w_k = d_k \nu_k^{(3)}$, with*

$$\nu_k^{(3)} = 4(2 - \lambda' \mathbf{x}_k q_k)^{-2}, \quad \lambda' = \mathbf{t}_x \cdot \left(\sum_{k \in s} \frac{w_k^2 q_k^2 \mathbf{x}_k \mathbf{x}_k'}{2 q_k (w_k + \sqrt{w_k d_k})} \right)^{-1}.$$

PROPOSITION 2. *The weights w_k, which satisfy the calibration equation $t_x = \hat{t}_x$ and minimize loss function L_6 are equal $w_k = d_k \nu_k^{(6)}$, with*

$$\nu_k^{(6)} = 1 + \left(\sum_{k=1}^{N} \mathbf{x}_k' - \sum_{k \in s} d_k \mathbf{x}_k' \right) \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k' q_k d_k^2 \right)^{-1} \mathbf{x}_k q_k d_k.$$

The proofs of Propositions 1 and 2 are based on the Lagrange technique.

Table 2. Auxiliary vector $\mathbf{x} = x$
True value of total: $t = 10652$
Sample size: $n = 20$

Loss function	Estimate of total	Estimated variance	Bias	MSE	cv	$\max_{1 \leqslant k \leqslant m} \lvert d_k - w_k \rvert$	$\frac{1}{m} \sum_{k=1}^{m} \lvert d_k - w_k \rvert$
Coefficient of correlation 0.8							
L_1	10597	518934	−54.64	521920	0.068	0.3321	0.1839
L_3	10706	801608	54.37	804564	0.084	0.0690	0.0479
L_6	10568	508563	−83.72	515572	0.067	0.3371	0.1844
L_7	10723	764058	71.51	769172	0.082	0.0690	0.0481
Coefficient of correlation 0.6							
L_1	10580	920000	−71.04	925046	0.091	0.3408	0.1836
L_3	10694	1018372	42.26	1020158	0.094	0.0705	0.0489
L_6	10587	904126	−64.74	908317	0.090	0.3426	0.1830
L_7	10706	1006671	54.80	1009675	0.094	0.0709	0.0488
Coefficient of correlation 0.4							
L_1	10631	1173986	−20.35	1174400	0.102	0.3255	0.1789
L_3	10731	1230813	79.84	1237188	0.103	0.0700	0.0488
L_6	10613	1146030	−38.38	1147503	0.101	0.3233	0.1776
L_7	10716	1193852	64.29	1197986	0.102	0.0693	0.0484
Coefficient of correlation 0.2							
L_1	10673	1421464	21.61	1421932	0.112	0.3204	0.1799
L_3	10778	1328522	126.50	1344524	0.107	0.0671	0.0483
L_6	10660	1462771	8.12	1462837	0.113	0.3214	0.1793
L_7	10725	1290466	73.18	1295821	0.106	0.0670	0.0480

The approximate variance of the presented calibrated estimators can be found by the Taylor linearization method. As an example we will present the approximate variance of the calibrated estimator of the total for the case of the loss function L_6.

PROPOSITION 3. *The approximate variance of the estimator*

$$\hat{t}_{yw} = \sum_{k \in \mathbf{i}} d_k v_k^{(6)} y_k$$

is

$$var(\hat{t}_w) \approx \sum_{k,l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l},$$

Here $e_k = y_k - \mathbf{x}_k' \mathbf{t}_3{}^{-1} \mathbf{t}_2$, $\mathbf{t}_2 = \sum_{k=1}^{N} d_k \mathbf{x}_k q_k y_k$, $\mathbf{t}_3 = \sum_{k=1}^{N} d_k \mathbf{x}_k \mathbf{x}_k' q_k$.

5. Simulation results

An artificial population of size $N = 100$ consisting of two strata of equal size is taken for the simulation. The sample size $n = 20$. One and two dimensional auxiliary variables x and $(1, x)$ are used: $\mathbf{x}_k = x_k$ and $\mathbf{x}_k = (1, x_k)$, $k = 1, \dots, N$ (Tables 1, 2).

Different variables $\mathbf{x}$, having a different correlation with the study variable y are examined. The calibrated estimators with the respective loss functions L_1, L_3, L_6, and L_7 are compared. $m = 1000$ stratified simple random samples are taken. The bias, variance, mean square error, and the coefficient of variation of the estimators of the total are estimated in the cases mentioned above. Two variability characteristics of the calibrated weights are calculated:

$$\max_{1 \le k \le m} |d_k - w_k| \quad \text{and} \quad \frac{1}{m} \sum_{k=1}^{m} |d_k - w_k|.$$

The simulation results show, that the variability of the weights w_k is smaller in the case of loss functions L_3 and L_7. The calibrated estimator under the loss function L_6 seems to be more stable.

References

1. J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382 (1992).
2. S. Lundström, *Calibration as Standard Method for Treatment of Nonresponse*, Doctoral dissertation, Stockholm University (1997).
3. S. Lundström, C.-E. Särndal, Calibration as standard method for treatment of nonresponse, *Journal of Official Statistics*, **15**(2), 305–327 (1999).
4. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York (1992).
5. A. Plikusas, Calibrated estimators of the ratio, *Lith. Math. J.*, **41**, 457–462 (2001).
6. A. Plikusas, Calibrated weights for the estimators of the ratio, *Lith. Math. J.*, **43**, 543–547 (2003).

REZIUMĖ

A. Plikusas, D. Pumputis. Kalibruoti sumų įvertiniai, esant skirtingoms nuostolių funkcijoms

Straipsnyje randami kalibruotų baigtinės populiacijos sumų įvertiniai, esant skirtingoms nuostolių funkcijoms. Pateikiama apytikslė kalibruoto sumos įvertinio dispersija nuostolių funkcijos L_6, pasiūlytos darbe [6], atveju. Pateikiami empiriniai kalibracinių svorių, esant skirtingoms nuostolių funkcijoms, palyginimai.