Liet. matem. rink., T. 42, spec. nr., 559–564

Calibrated estimators of the ratio with multidimensional auxiliary information

Aleksandras PLIKUSAS (MII), Alina PETRIKAITĖ (MII)
e-mail: plikusas@ktl.mii.lt, alina_26@yahoo.com

1. Introduction and notation

Calibrated estimators are widely used in survey sampling. The calibrated estimator of a total was presented by Devile and Särndal [1]. The calibrated estimator of a ratio was introduced by Plikusas [5]. In the case of significant correlation between study and auxiliary variables, the variance of a calibrated estimator is lower. The calibrated estimators can also be efficiently used in the presence of nonresponse (see Lundström [2], Lundström and Särndal [3]).

A finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements is considered. Let y and z be two study variables defined on the population $\mathcal{U}$ and taking values $\{y_1, y_2, \ldots, y_N\}$ and $\{z_1, z_2, \ldots, z_N\}$, respectively. Denote by t_y and t_z unknown population totals of y and z:

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

We are interested in the estimation of the ratio of two totals $R = t_y/t_z$. There are many examples in survey statistics where the ratio is the main parameter to be estimated. The survey of wages and salaries is an important example of such kind. Auxiliary information can be taken from the previous complete surveys and various enterprise registers.

2. Calibrated estimator of the ratio

Denote by $\widehat{t}_y$ and $\widehat{t}_z$ the Horwitz–Thompson estimators of the totals t_y and t_z:

$$\widehat{t}_y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k, \quad \widehat{t}_z = \sum_{k \in s} \frac{z_k}{\pi_k} = \sum_{k \in s} d_k z_k.$$

Here $s \subset \{1, 2, \ldots, N\}$ is a set of indices of a random sample from the population $\mathcal{U}$; π_k denote the inclusion probability of the element u_k into the sample; d_k is usually called the design weight of the element u_k, $k = 1, 2, \ldots, N$.

Given estimators $\widehat{t}_y$ and $\widehat{t}_z$ one can use a straightforward estimator of the ratio $\widehat{R} = \widehat{t}_y/\widehat{t}_z$. We offer another ratio estimator that uses some known auxiliary information.

Suppose that a vector valued variables $\mathbf{x}_y$ and $\mathbf{x}_z$ are known for the variables y and z correspondingly:

$$\mathbf{x}_{yk} = (x_{yk1}, \ldots, x_{ykm}), \quad \mathbf{x}_{zk} = (x_{zk1}, \ldots, x_{zkm}), \quad k = 1, 2, \ldots, N.$$

Denote $\mathbb{R}_0 = (R_{01}, \ldots, R_{0m})$, where

$$R_{0j} = \frac{\sum_{k=1}^{N} x_{ykj}}{\sum_{k=1}^{N} x_{zkj}}, \quad j = 1, \ldots, m$$

are known population totals.

We propose a calibrated estimator of the ratio

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

with weights $w_k, k \in s$, called calibration weights. Calibration weights have to satisfy two requirements. The first one is the calibration equation:

$$\mathbb{R}_{0w} = \mathbb{R}_0, \tag{1}$$

with $\mathbb{R}_{0w} = (R_{0w1}, \ldots, R_{0wm})$ and

$$R_{0wj} = \frac{\sum_{k \in s} w_k x_{ykj}}{\sum_{k \in s} w_k x_{zkj}}, \; j = 1, \ldots, m.$$

The second requirement is that calibration weights be as close to the initial design weights as possible. We use the usual loss function to measure the distance between the weights:

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

Here q_k, $k = 1, 2, \ldots N$, are some individual positive weights. We are free to choose q_k and by choosing q_k we can get different estimators of R. Write $\mathbf{x}'$ for the transposed column vector $\mathbf{x}$.

PROPOSITION 1. The calibrated weights satisfying (1) and minimizing (2) are

$$w_k = d_k(1 + q_k \mathbf{a}_k' \widehat{\mathbf{g}}),$$

here $\mathbf{a}_k' = (a_{k1}, \ldots, a_{km})$, $a_{kj} = x_{ykj} - R_{0j} x_{zkj}$, $j = 1, \ldots, m$, and

$$\widehat{\mathbf{g}} = -\widehat{\mathbf{T}}^{-1} \sum_{k \in s} d_k \mathbf{a}_k, \quad \widehat{\mathbf{T}} = \sum_{k \in s} d_k q_k \mathbf{a}_k \mathbf{a}_k'.$$

Proof. Define the Lagrange function

$$l = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k q_k} - \boldsymbol{\lambda}'(\mathbb{R}_{0w} - \mathbb{R}_0), \quad \boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_m).$$

The derivatives $\dfrac{\partial l}{\partial w_k}$ are equal to zero when

$$w_k = d_k(1 + q_k \mathbf{a}_k' \boldsymbol{\Lambda}), \quad \boldsymbol{\Lambda}' = (\Lambda_1, \ldots, \Lambda_m), \tag{3}$$

$$\Lambda_j = \frac{\lambda_j}{2 \sum_{k \in s} w_k x_{zkj}}, \quad j = 1, \ldots, m.$$

It follows from (3), that

$$\sum_{k \in s} w_k x_{ykj} = \sum_{k \in s} d_k x_{ykj} + \sum_{i=1}^{m} \Lambda_i \sum_{k \in s} x_{ykj} a_{ki} d_k q_k, \quad j = 1, \ldots, m, \tag{4}$$

and

$$\sum_{k \in s} w_k x_{zkj} = \sum_{k \in s} d_k x_{zkj} + \sum_{i=1}^{m} \Lambda_i \sum_{k \in s} x_{zkj} a_{ki} d_k q_k, \quad j = 1, \ldots, m. \tag{5}$$

Combining (4) and (5), we can find that

$$\widehat{\mathbf{T}} \boldsymbol{\Lambda} = -\sum_{k \in s} d_k \mathbf{a}_k.$$

This completes the proof.

Note that the calibrated weights w_k are random: they depend on the sample s. So the calibrated estimator of the ratio $\widehat{R}_w$ as well as known calibrated estimator of the total is more complicated than straightforward estimator $\widehat{R}$.

An approximate variance of the estimator $\widehat{R}_w$ can be found using Taylor's linearization method. The variance of the linear part of the Taylor expansion is usually used as an approximate variance of complicated estimators. We need some additional notation for the expression of the approximate variance. Denote by $\mathbf{T}$ the $m \times m$ matrix

$$\mathbf{T} = \sum_{k=1}^{N} d_k q_k \mathbf{a}_k \mathbf{a}_k',$$

and by T_{ij}, $i, j = 1, \ldots, m$, the algebraic adjunct of the element t_{ij} of matrix $\mathbf{T}$. Write: $\mathbf{T}_i' = (T_{i1}, \ldots, T_{im})$,

$$\mathbf{T}_y = \sum_{k=1}^{N} d_k q_k y_k \mathbf{a}_k', \quad \mathbf{T}_z = \sum_{k=1}^{N} d_k q_k z_k \mathbf{a}_k'.$$

Now we can formulate

PROPOSITION 2. The approximate variance of the estimator $\widehat{R}_w$ is

$$Avar(\widehat{R}_w) \approx \frac{1}{t_z^2} \sum_{k,l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}.$$

Here

$$e_k = y_k - Rz_k + \sum_{i=1}^{m} b_i a_{ki}, \quad b_i = (R\mathbf{T}_z - \mathbf{T}_y) \frac{\mathbf{T}_i}{|\mathbf{T}|}.$$

π_{kl} is the inclusion probability of the pair of elements u_k and u_l into the sample.

We omit the proof which is completely technical.

3. Some properties of the calibrated estimator of the ratio

Let us compare the variance of the calibrated estimator of the ratio with the variance of some other possible estimators of the ratio. Let variables x_y: $x_{y1}, x_{y2}, \ldots, x_{yN}$, x_z: $x_{z1}, x_{z2}, \ldots, x_{zN}$ serve as auxiliary variables for the study variables y and z, respectively,

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

One can take the ratio estimators of totals t_y and t_z and consider the following ratio estimator of ratio

$$\widehat{R}_{rat} = \frac{\frac{\widehat{t}_y}{t_{xy}} t_{xy}}{\frac{\widehat{t}_z}{t_{xz}} t_{xz}} = \frac{t_{xy}}{t_{xz}} \cdot \frac{\widehat{t}_y \widehat{t}_{xz}}{\widehat{t}_z \widehat{t}_{xy}} = R_0 \frac{\widehat{t}_y \widehat{t}_{xz}}{\widehat{t}_z \widehat{t}_{xy}}.$$

The ratio

$$R_0 = \frac{t_{xy}}{t_{xz}}$$

is supposed to be known. The approximate variance of such an estimator is

$$Avar(\widehat{R}_{rat}) = \frac{1}{t_z^2} Var\left((\widehat{t}_y - R\widehat{t}_z) - \frac{t_y}{t_{xy}}(\widehat{t}_{xy} - R_0 \widehat{t}_{xz})\right).$$

Another possibility is to take the regression estimator of totals and

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

where

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

$$\widehat{B}_z = \frac{\sum_{k\in s} d_k(z_k - \widehat{\mu}_z)(x_{zk} - \widehat{\mu}_{xz})}{\sum_{k\in s} d_k(x_{zk} - \widehat{\mu}_{xz})^2}.$$

The approximate variance of the estimator $\widehat{R}_{reg}$ is (see, for example, [4])

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

where

$$B_y = \frac{\sum_{k=1}^N (y_k - \mu_y)(x_{yk} - \mu_{xy})}{\sum_{k=1}^N (x_{yk} - \mu_{xy})^2},$$

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

Now we formulate some properties of the estimators of the ratio.

1. Approximate variance of the calibrated estimator of the ratio is not higher than the approximate variance of a simple ratio estimator for any sample design:

$$AVar(\widehat{R}_w) \leqslant AVar(\widehat{R}).$$

2. Under the condition

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

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

$$AVar(\widehat{R}_{rat}) \leqslant AVar(\widehat{R})$$

for any sample design. Here ρ denote the correlation coefficient.

3. Approximate variance of the calibrated estimator of the ratio is not higher than the approximate variance of the ratio estimator of a ratio for any sample design:

$$AVar(\widehat{R}_w) \leqslant AVar(\widehat{R}_{rat}).$$

Table 1

n	$\widehat{cv}(\widehat{R})$	$\widehat{cv}(\widehat{R}_w)$	$\widehat{cv}(\widehat{R}_{rat})$
15	0.21	0.17	0.17
16	0.19	0.15	0.16
17	0.19	0.15	0.15
18	0.18	0.13	0.14
19	0.16	0.13	0.13
20	0.15	0.12	0.12
21	0.14	0.11	0.12
22	0.13	0.11	0.11

4. Some simulation results

Some preliminary simulation was performed for the small population with $N = 32$. The data were taken from the Lithuanian survey on wages and salaries. The historical data (from the previous complete survey) were used as auxiliary information. A simple random sample was examined. Usually a stratified sample is used in such enterprise surveys. That is a motivation for the small population size. The Table 1 presents the coefficients of variation for sample sizes from 15 to 22. It should be mentioned that the calibrated estimator is more efficient (i.e., has smaller variance) for larger populations, even in the case where study variables are not strongly correlated with the auxiliary information ($\rho = 0.3 - 0.5$).

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).

Kalibruoti santykio įvertiniai, esant daugiamatei papildomai informacijai

A. Plikusas, A. Petrikaitė

Straipsnyje nagrinėjamas kalibruotas dviejų sumų santykio įvertinys. Kalibruoti įvertiniai paprastai turi mažesnes dispersijas, juos galima taikyti, esant neatsakymams į apklausas. Rasta pasiūlyto įvertinio apytikslė dispersija, kuri vienamačiu atveju palyginta su kai kurių kitų jau žinomų įvertinių dispersijomis. Gauti preliminarūs modeliavimo rezultatai mažai populiacijai.