Liet. matem. rink., T. 41, spec. nr., 457–462


# Calibrated estimators of the ratio

Aleksandras PLIKUSAS (MII)
*e-mail: plikusas@ktl.mii.lt*

## 1. Introduction

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of $N$ elements and two study variables $y$ and $z$ taking values $\{y_1, y_2, \ldots, y_N\}$ and $\{z_1, z_2, \ldots, z_N\}$, respectively. Let $t_y$ and $t_z$ denote 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 various statistical methods for improving estimators using auxiliary information. The calibration method is one of them. The idea of calibration of the estimators of totals was presented in 1992 by Deville and Särndal [1]. A calibration estimator uses calibrated weights, which are as close as possible, according to a given distance measure, to the original sample design weights and satisfies some calibration equations. In this paper, the calibrated estimator of the ratio is introduced. Calibrated estimators can be used in the presence of nonresponse, e.g., [2], [3].

## 2. Calibrated estimators 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 $\mathcal{U}$; $\pi_k$, $k = 1, 2, \ldots, N$, denote the inclusion probability of the element $u_k$ into the sample; $d_k$, $k = 1, 2, \ldots, N$, is usually called as design weight of the element $u_k$.

Suppose, the variables of auxiliary information $x_y$ and $x_z$ are available for the study variables $y$ and $z$ correspondingly. It means that we know population values $x_{y1}, x_{y2}, \ldots, x_{yN}$ and $x_{z1}, x_{z2}, \ldots, x_{zN}$, where $x_y$ serves as auxiliary information for the study variable $y$ and $x_z$ – for the study variable $z$.This auxiliary information may be

known from the previous census, administrative data, other sources. So, we assume, that the population totals

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

are known.

One possibility of constructing calibrated estimator of the ratio $R$ is as follows. We can construct calibrated estimators of the totals $t_y$ and $t_z$ using auxiliary vector $\mathbf{x}_k = (x_{yk}, x_{zk})$ for the element $u_k$ as in Deville and Särndal (1992). The ratio of these estimators of totals can be taken as an estimator of the ratio $R$.

Let us consider a different calibrated estimator. Suppose that the variable $y$ is associated with the auxiliary variable $x_y$, and $z$ is associated with $x_z$. Denote the known ratio by

$$R_0 = \frac{\sum_{k=1}^{N} x_{yk}}{\sum_{k=1}^{N} x_{zk}},$$

and define the following calibration equation

$$R_0 = \frac{\sum_{k\in s} w_k\, x_{yk}}{\sum_{k\in s} w_k\, x_{zk}}. \tag{1}$$

We use the standard loss function

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

to measure the difference between the design weights $d_k$ and calibrated weights $w_k$. 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 $R$.

PROPOSITION 1. *The calibrated estimator, whose weights $w_k$ satisfy* (1) *is given by*

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

*where*

$$w_k = d_k\left(1 - q_k \frac{\sum_{k\in s} d_k(x_{yk} - R_0 x_{zk})}{\sum_{k\in s} d_k q_k (x_{yk} - R_0 x_{zk})^2}\,(x_{yk} - R_0 x_{zk})\right), \quad k \in s. \tag{3}$$

*Proof.* Define the Lagrange function

$$L = \sum_{k\in s} \frac{(w_k - d_k)^2}{d_k\, q_k} - \lambda\left(\frac{\sum_{k\in s} w_k\, x_{yk}}{\sum_{k\in s} w_k\, x_{zk},} - R_0\right).$$

The derivatives of $L$ are

$$\frac{\partial L}{\partial w_k} = \frac{2w_k}{d_k q_k} - \frac{2}{q_k} - \lambda \frac{x_{yk} - R_0 x_{zk}}{\sum_{k\in s} w_k x_{zk}}.$$

These derivatives equal zero when

$$w_k = d_k + \frac{\lambda}{2} d_k q_k \frac{x_{yk} - R_0 x_{zk}}{\sum_{k\in s} w_k x_{zk}}. \tag{4}$$

From (4) we can get

$$\sum_{k\in s} w_k x_{yk} = \sum_{k\in s} d_k x_{yk} + \frac{\lambda}{2} \frac{\sum_{k\in s} x_{yk}(x_{yk} - R_0 x_{zk}) d_k q_k}{\sum_{k\in s} w_k x_{zk}}, \tag{5}$$

and

$$\sum_{k\in s} w_k x_{zk} = \sum_{k\in s} d_k x_{zk} + \frac{\lambda}{2} \frac{\sum_{k\in s} x_{zk}(x_{yk} - R_0 x_{zk}) d_k q_k}{\sum_{k\in s} w_k x_{zk}}. \tag{6}$$

Summing (5) and (6) multiplied by $R_0$, we get

$$\lambda = -2 \frac{\sum_{k\in s} d_k (x_{yk} - R_0 x_{zk}) \sum_{k\in s} w_k x_{zk}}{\sum_{k\in s} d_k q_k (x_{yk} - R_0 x_{zk})^2}. \tag{7}$$

Inserting $\lambda$ from (7) into (4) the expressions (3) are obtained.

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

$$var(\widehat{R}_w) \approx \frac{1}{t_z^2} \sum_{k,l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{8}$$

*Here*

$$e_k = y_k - R z_k - R_1 (x_{yk} - R_0 x_{zk}),$$

$$R_1 = \frac{t_y \sum_{k=1}^{N} q_k (x_{yk} - R_0 x_{zk}) - t_z \sum_{k=1}^{N} q_k (x_{yk} - R_0 x_{zk}) y_k}{t_z \sum_{k=1}^{N} q_k (x_{yk} - R_0 x_{zk})^2},$$

*$\pi_{kl}$ is the inclusion probability of the pair of elements $u_k$ and $u_l$ into the sample.*

*Proof.* We use Taylor linearization technique. The estimator $\widehat{R}_w$ can be written in the following form

$$\widehat{R}_w = \frac{\widehat{t}_y \widehat{t}_1 - \widehat{t}_2 \widehat{t}_3}{\widehat{t}_z \widehat{t}_1 - \widehat{t}_2 \widehat{t}_4},$$

here

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

Taking the derivatives of $\widehat{R}_w$ by $\widehat{t}_y, \widehat{t}_z, \widehat{t}_1, \widehat{t}_2, \widehat{t}_3, \widehat{t}_4$ at the point $(t_y, t_z, t_1, t_2, t_3, t_4)$ we can derive the linearized estimator

$$\widehat{R}_{wl} = R + \frac{1}{t_z}\big((\widehat{t}_y - t_y) + R_1(\widehat{t}_2 - t_2) - R(\widehat{t}_z - t_z)\big),$$

where $R_1 = \frac{R}{t_1}(t_4 - t_3)$. A reasonable choice of the point $t_1, t_2, t_3, t_4$ is

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

Then the variance of the linearized estimator is equal to

$$var(\widehat{R}_{wl}) = \frac{1}{t_z^2} var\Big(\sum_{k\in s} d_k e_k\Big),$$

where $e_k = y_k + R_1(x_{yk} - R_0 x_{zk}) - R z_k$. Using the Result 2.8.1 from [4], we find that

$$var(R_{wl}) = \frac{1}{t_z^2} \sum_{k,l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}.$$

The proof is complete.

PROPOSITION 3. *As the variance estimator of the calibrated estimator $\widehat{R}_w$ we can take*

$$\widehat{var}(\widehat{R}_w) = \frac{1}{\widehat{t}_z^2} \sum_{k,l\in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{\widehat{e}_k}{\pi_k} \frac{\widehat{e}_l}{\pi_l}. \tag{9}$$

*Here*

$$\widehat{e}_k = y_k - \widehat{R}_w z_k - \widehat{R}_1 (x_{yk} - R_0 x_{zk}),$$
$$\widehat{R}_1 = \frac{\widehat{t}_y \sum_{k=1}^{N} q_k (x_{yk} - R_0 x_{zk}) - \widehat{t}_z \sum_{k\in s} d_k q_k (x_{yk} - R_0 x_{zk}) y_k}{\widehat{t}_z \sum_{k=1}^{N} q_k (x_{yk} - R_0 x_{zk})^2}.$$

Let us comment the variance estimator proposed in Proposition 3. Using result 2.8.1 of [4], the unbiased estimator of the variance (8) is

$$\widehat{var}\,\widehat{R}_{wl} = \sum_{k,l\in s}\left(1-\frac{\pi_k\pi_l}{\pi_{kl}}\right)\frac{e_k}{\pi_k}\frac{e_l}{\pi_l}.$$

The values $e_k$ contain the unknown parameter $R_1$. We suggest to use

$$\widehat{R}_1 = \frac{R_0}{\widehat{t}_1}(\widehat{t}_4 - \widehat{t}_3),$$

as the estimator of $R_1$. Inserting $\widehat{R}_1$ instead of $R_1$ in the expressions of $e_k$, we derive estimator (9).

## 3. The case of nonresponse

Nonresponse by some selected elements of a sample occurs in almost every survey. It is often the case, that responding elements have a different distribution as responding ones. As a result some bias occurs if the nonrespondence is ignored. The calibration may be considered as a unified approach for reducing nonresponse bias, using auxiliary information.

Let us suppose, that some sampled elements do not respond and denote by $r$, $r \subset s$, the set of responded elements. Define a new calibration equation:

$$R_0 = \frac{\sum_{k\in r} w_k\, x_{yk}}{\sum_{k\in r} w_k\, x_{zk}}. \tag{10}$$

The summation in (10) is taken over the set of responded elements. The calibration equation requires that new weights $w_k$ estimates the auxiliary ratio exactly. This auxiliary ratio may be taken from the previous census, administrative data or other sources.

PROPOSITION 4. *The calibrated estimator, whose weights $w_k$ satisfy* (10) *and minimize the loss function*

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

*is given by* $\widehat{R}_r = \dfrac{\sum_{k\in r} w_k\, y_k}{\sum_{k\in r} w_k\, z_k}$, *where*

$$w_k = d_k\left(1 - q_k\frac{\sum_{k\in r} d_k(x_{yk}-R_r x_{zk})}{\sum_{k\in r} d_k q_k (x_{yk}-R_r x_{zk})^2}(x_{yk}-R_r x_{zk})\right), \quad k\in r.$$

*Proof.* The proof of this statement is similar to that of Proposition 1.

The weights $w_k$ in the case of nonresponse differ from the calibrated weights derived in Proposition 1.

## 4. Other loss functions

Another possibility of constructing new estimators is the choice of a different loss function. We produced calibrated weights so far, using the loss function with summands $(w_k - d_k)^2/(d_k q_k)$. Some other loss functions were considered by J.-C. Deville and C.-E. Särndal in [1]. Unfortunately, the calibration of the ratio estimator does not lead to an explicit solution. But in some cases, approximate solution be the same. For example, in the case of the loss function

$$L = -\frac{d_k}{q_k} \log \frac{w_k}{d_k} + \frac{1}{q_k}(w_k - d_k),$$

we can get the equation

$$w_k = \frac{d_k}{q_k} - b \frac{w_k(x_{yk} - R_0 x_{zk})}{\sum_{k \in s} w_k (x_{yk} - R_0 x_{zk})^2}, \quad k \in s, \tag{11}$$

where

$$b = \sum_{k \in s} \frac{d_k}{q_k(x_{yk} - R_0 x_{zk})}.$$

Equation (11) may be used in the iterative procedure for the approximate solution.

## 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).

# Kalibruoti santykio įvertiniai

A. Plikusas

Straipsnyje pateikiamas kalibruotas dviejų sumų santykio įvertinys. Kalibruoti įvertinai – tai įvertiniai, kuriuose panaudojama papildoma informacija, norint gauti tikslesnius vertinamų parametrų įverčius. Yra žinoma, kad atskirais atvejais kalibruoti įvertiniai sutampa su santykiniais įvertiniais, kurie yra tikslesni (turi mažesnes dispersijas), jei tiriamas kintamasis yra pakankamai gerai koreliuotas su papildomu žinomu kintamuoju. Darbe pateikta siūlomo įvertinio apytikslė dispersija ir tos dispersijos įvertinys.