Liet. matem. rink., T. 43, spec. nr., 543–547

Calibrated weights for the estimators of the ratio

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

1. Introduction

There are various statistical methods for improving estimators, using auxiliary information. The calibration method is one of them. The growing power of calculation abilities stimulates consideration of estimators, based on the exténsive use of auxiliary information. This kind of estimators can be used in official statistics. The calibration technique for the estimation of population totals was presented in [1]. In [1] and [2] the calibrated estimators of totals for estimating in the presence of nonresponse are considered. In this paper, a nonlinear calibration problem is formulated. Some calibrated estimators of the ratio of totals are proposed.

2. Calibration problem

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements and q population variables $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(q)}$, where $\mathbf{y}^{(j)}$ is taking values $y_{j1}, \ldots, y_{jN}$.

Let θ be a parameter to be estimated and it is expressed as a function of q population totals $\mathbf{t} = (t_1, \ldots, t_q)$:

$$\theta = f(\mathbf{t}) = f(t_1, \ldots, t_q), \quad t_j = \sum_{k=1}^{N} y_{jk}.$$

Suppose that for each population element k, $k = 1, \ldots, N$, and the variable $\mathbf{y}^{(j)}$, the vector of auxiliary information $\mathbf{x}_k^{(j)} = (x_{k1}^{(j)}, \ldots, x_{k,m_j}^{(j)})$ is known. It means that, for different study variables $\mathbf{y}^{(j)}$, $j = 1, \ldots, q$, we may use a different $N \times m_j$ matrix of auxiliary information.

Denote by $\hat{\mathbf{t}} = (\hat{t}_1, \ldots, \hat{t}_q)$ the Horvitz–Thompson estimator of the totals $\mathbf{t} = (t_1, \ldots, t_q)$:

$$\hat{t}_j = \sum_{k \in s} \frac{y_{jk}}{\pi_k} = \sum_{k \in s} d_k y_{jk}, \quad j = 1, \ldots, q.$$

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 as a design weight of the element u_k, $k = 1, 2, \ldots, N$.

Write

$$\mathbf{t}_x^{(j)} = \sum_{k=1}^{N} \mathbf{x}_k^{(j)}.$$

Let us calibrate the estimators $\hat{t}_j$, i.e., find new weights $w_k^{(j)}$, $k \in s$, which:

1) for some (possibly vector-valued) function g

$$g(\hat{\mathbf{t}}_x^{(1)}, \dots, \hat{\mathbf{t}}_x^{(q)}) = g\left(\mathbf{t}_x^{(1)}, \dots, \mathbf{t}_x^{(q)}\right)$$

with

$$\hat{\mathbf{t}}_x^{(j)} = \sum_{k\in s} w_k^{(j)} \mathbf{x}_k^{(j)}, \quad j = 1, \dots, q.$$

2) the distance between $w_k^{(j)}$ and the design weights d_k is minimized by some distance measure.

Now the calibrated estimators of totals t_j can be defined as

$$\hat{t}_j(cal) = \sum_{k\in s} w_k^{(j)} y_{jk}, \quad j = 1, \dots, q,$$

and the calibrated estimator of θ as

$$\hat{\theta}(cal) = f\left(\hat{t}_1(cal), \dots, \hat{t}_q(cal)\right).$$

Note, that different auxiliary information can be used for different study variables and different weight systems for the estimation of different totals t_j. This feature enables us to select separately the auxiliary information for different study variables.

3. Examples of distance measures

Let us introduce free additional weights q_k, $k = 1, \dots, N$. One can modify estimators by choosing q_k. Otherwise, we can put $q_k = 1$ for all k. The following distance measures 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 measures $L_1 - L_5$ are mentioned in [1].

4. Calibrated estimators of the ratio

Consider 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 $R = t_y/t_z$.

Suppose, the variables of auxiliary information $x^{(y)}$ and $x^{(z)}$ are available for the study variables y and z, respectively. It means that we know the population values $x_1^{(y)}, x_2^{(y)}, \ldots, x_N^{(y)}$ and $x_1^{(z)}, x_2^{(z)}, \ldots, x_N^{(z)}$, where $x^{(y)}$ serves as auxiliary information for the study variable y and $x^{(z)}$ – for the study variable z. In official statistics this auxiliary information may be known from the previous census, administrative data, other sources.

So, we assume that the population totals

$$t_x^{(y)} = \sum_{k=1}^{N} x_k^{(y)}, \quad t_x^{(z)} = \sum_{k=1}^{N} x_k^{(z)}$$

are known. It means we know the ratio $R_0 = t_x^{(y)}/t_x^{(z)}$. Let us formulate a calibration problem in this particular case: find a system of calibrated weights $w_k^{(1)}$, $w_k^{(2)}$, $k \in s$, which satisfy calibration equation:

$$R_0 = \frac{\sum_{k\in s} w_k^{(1)} x_k^{(y)}}{\sum_{k\in s} w_k^{(2)} x_k^{(z)}}, \tag{1}$$

and minimize the loss function

$$L^2(w,d) = \sum_{k\in s} \frac{(w_k^{(1)} - d_k)^2}{d_k q_k} + \sum_{k\in s} \frac{(w_k^{(2)} - d_k)^2}{d_k q_k}. \tag{2}$$

The respective ratio estimator will be

$$\widehat{R}_w^{(1)} = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k}. \tag{3}$$

The solution of this calibration problem is as follows.

PROPOSITION 1. The weights $w_k^{(1)}$, $w_k^{(2)}$, $k \in s$ of the calibrated estimator (3) which satisfy (1) and minimize (2) are given by

$$w_k^{(1)} = d_k\left(1 - q_k \frac{\sum_{k\in s} d_k (x_k^{(y)} - R_0 x_k^{(z)})}{\sum_{k\in s} d_k q_k ((x_k^{(y)})^2 + R_0^2 (x_k^{(z)})^2)} x_k^{(y)}\right), \quad k \in s.$$

$$w_k^{(2)} = d_k\left(1 + q_k \frac{\sum_{k\in s} d_k(x_k^{(y)} - R_0 x_k^{(z)})}{\sum_{k\in s} d_k q_k((x_k^{(y)})^2 + R_0^2(x_k^{(z)})^2)} R_0 x_k^{(z)}\right), \quad k \in s.$$

The proof of this proposition is similar to that of the special case when $w_k^{(1)} = w_k^{(2)}$, and can be found in [5].

The distance measure L_1 alone is being used in practice, even calibrating estimators of population totals. Let us take the distance measure L_6 and $w_k^{(1)} = w_k^{(2)}$, $k \in s$, for simplicity. Then the calibrated estimator of the ratio R is of the form

$$\widehat{R}_w^{(1)} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k z_k}.$$

Denote

$$\widehat{R}_0 = \frac{\sum_{k\in s} w_k x_k^{(y)}}{\sum_{k\in s} w_k x_k^{(z)}}.$$

PROPOSITION 2. The weights w_k, $k \in s$ of the calibrated estimator $R_w^{(1)}$ which satisfy the calibration equation $R_0 = \widehat{R}_0$ and minimize L_6 are given by

$$w_k = d_k - \frac{q_k d_k^2 r_k}{\sum_{k\in s} q_k d_k^2 r_k^2} \sum_{k\in s} d_k r_k,$$

here $r_k = x_k^{(y)} - R_0 x_k^{(z)}$.

Proof. Define the Lagrange function

$$l = \sum_{k\in s} \frac{1}{q_k}\left(\frac{w_k}{d_k} - 1\right)^2 - \lambda(\widehat{R}_0 - R_0).$$

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

$$w_k = d_k + \frac{\lambda}{2\hat{t}_x^{(z)}} q_k d_k^2 \hat{r}_k, \tag{4}$$

where $\hat{r}_k = x_k^{(y)} - \widehat{R}_0 x_k^{(z)}$, $\hat{t}_x^{(z)} = \sum_{k\in s} w_k x_k^{(z)}$. Summing (4) multiplied by $\hat{r}_k$, we can find

$$\lambda = -\frac{2\hat{t}_x^{(z)} \sum_{k\in s} d_k r_k}{\sum_{k\in s} q_k d_k^2 r_k^2}. \tag{5}$$

The proposition follows from (4), (5) and the calibration equation $\widehat{R}_0 = R_0$.

The weights w_k can be negative. This feature is unacceptable in the practice of the official statistics. The generalized calibration problem for the ratio does not have an explicit solution for the distance measures $L_2 - L_5$ and L_7. In the case L_2 and L_3 the approximate solution is the same as for L_1.

Let us consider the distance measure L_7 and the case $w_k^{(1)} = w_k^{(2)}$.

PROPOSITION 3. The weights w_k, $k \in s$, that satisfy the calibration equation $R_0 = \widehat{R}_0$ and minimize L_7, satisfy the equation

$$w_k = \frac{d_k(\widehat{t}_x^{(z)})^2}{(1 - bq_kd_kr_k)^2}, \quad b = \widehat{t}_x^{(z)} \frac{\sum_{k\in s}(q_k^2 d_k r_k)^{-1}\left(\frac{w_k}{d_k} - 1\right)}{2\sum_{k\in s} w_k (q_k d_k)^{-1}}. \tag{6}$$

Equations (6) can be used for the iterative procedure. At the initial step, one can take $w_k = d_k$ on the right hand side. The attractive property of this system of weights is positivity.

We can construct many different calibrated estimators of the ratio, using different calibration equations and different distance measures. The simulation study is needed to define preferences among many possible estimators of the ratio. The approximate variances of these estimators can be found using the linearization technique.

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 Methodfor 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, *Lithuanian Math. J.*, **41**, 457–462 (2001).

Santykio įvertinių kalibruoti svoriai

A. Plikusas

Straipsnyje suformuluotas apibendrintas netiesinės kalibracijos uždavinys. Pateikti dviejų sumų santykio įvertinio svorių kalibravimo pavyzdžiai. Pasiūlytos kelios atstumo funkcijos ir atitinkamos svorių sistemos.