LMD konf. darbai, 204–208
1997, Vilnius

On regression estimators of the ratio and their applications

G. Klimavičius, D. Krapavickaitė, A. Plikusas (IMI)

Sampling surveys usually have several study variables. Some auxiliary variables may be used at the estimation stage. For example, it may be data of the previous census or other data basis.

The paper is devoted to the estimation of variance of special statistics expressed as the ratio of products of population total estimators. The statistics of such type naturally arise when the ratio estimators are used for estimation of totals and we have to estimate the ratio of these totals. A real example of the case study is presented.

Let N be the size of finite population. Suppose we have $m+n$ study variables with the population values

$$t_{j1}, \dots, t_{jN}, \quad j = 1, \dots, m;$$

$$u_{j1}, \dots, u_{jN}, \quad j = 1, \dots, n.$$

So, we have $m+n$ population totals:

$$t_j = \sum_{k=1}^{N} t_{jk}, \quad j = 1, \dots, m,$$

$$u_j = \sum_{k=1}^{N} u_{jk}, \quad j = 1, \dots, n.$$

Some of these totals may be known, some of them may be unknown and have to be estimated from the sample.

We examine a statistic

$$\widehat{R} = \frac{\widehat{t_1} \cdot \ldots \cdot \widehat{t_m}}{\widehat{u_1} \cdot \ldots \cdot \widehat{u_n}}.$$

Here $\widehat{t_j}$, $j = 1, \dots, m$; $\widehat{u_j}$ $j = 1, \dots, n$, are Horvitz–Thompson estimators (π estimators) of the totals t_j and u_j:

$$\widehat{t_j} = \sum_{k \in S} \frac{t_{jk}}{\pi_k}, \quad j = 1, \dots, m,$$

$$\widehat{u}_j = \sum_{k \in S} \frac{u_{jk}}{\pi_k}, \quad j = 1, \ldots, n.$$

S denotes a set of sample elements; π_k, $k = 1, \ldots, N$, are inclusion probabilities of the k th population element into the sample. Let us denote by π_{kl}, $k, l = 1, \ldots, N$, an inclusion probability of a pair of elements k and l into the sample.

The objective is to estimate the variance of statistic $\widehat{R}$. It is known, that even in the case $m = n = 1$, only an approximate computable formula is available.

In order to calculate an approximate variance of $\widehat{R}$ we use standard Taylor's linearization procedure and obtain

$$\widehat{R} \approx \widehat{R}_L = R\left(1 + \sum_{j=1}^{m} \frac{\widehat{t}_j - t_j}{t_j} - \sum_{j=1}^{n} \frac{\widehat{u}_j - u_j}{t_j}\right).$$

Here

$$R = \frac{t_1 \cdot \ldots \cdot t_m}{u_1 \cdot \ldots \cdot u_n}.$$

The variance of $\widehat{R}_L$ can be calculated.

PROPOSITION 1. *The approximate variance of $\widehat{R}$ is obtained as*

$$\mathbf{D}\,\widehat{R} \approx \mathbf{D}\,\widehat{R}_L = R^2 \sum_{k,l=1}^{N} \Delta_{kl} \frac{z_k}{\pi_k} \frac{z_l}{\pi_l}, \tag{1}$$

where

$$z_k = \sum_{j=1}^{m} \frac{t_{jk}}{t_j} - \sum_{j=1}^{n} \frac{u_{jk}}{u_j}, \quad k = 1, \ldots, N,$$

and

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l.$$

PROPOSITION 2. *A variance estimator of statistic $\widehat{R}$ is given by*

$$\widehat{\mathbf{D}}\,\widehat{R} = \widehat{\mathbf{D}}\,\widehat{R}_L = \widehat{R}^2 \sum_{k,l \in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\widehat{z}_k}{\pi_k} \frac{\widehat{z}_l}{\pi_l}, \tag{2}$$

where

$$\widehat{z}_k = \sum_{j=1}^{m} \frac{t_{jk}}{\widehat{t}_j} - \sum_{j=1}^{n} \frac{u_{jk}}{\widehat{u}_j}, \quad k \in S. \tag{3}$$

If t_j or u_j are known for some indices j, the corresponding estimates $\widehat{t}_j$ or $\widehat{u}_j$ in (3) *are replaced by the known totals t_j, u_j.*

Formulas (1) and (2) may be derived using the result 5.6.2 of [1] and considering z_k, $k = 1, \ldots, N$, as new values of a new variable z.

Estimator (2) was used in real surveys. The case study is briefly presented below.

CASE STUDY LITHUANIAN ENTERPRISE SURVEY ON WAGES AND SALARIES

We examine some aspects of the monthly enterprise survey on wages and salaries. A sampling unit (SU) is a part of an individual company or institution engaged in a certain kind of economic activity. The sampling frame is based on the Business Register of Lithuanian Department of Statistics. The smallest domains of estimation are kinds of economic activity inside the two kinds of ownership – private and state. Let us call these domains the basic domains of estimation (BDE). So, every domain of estimation is a union of some BDEs. The BDEs are stratified according to the number of employees, using one-year old data of the previous census. The stratification boundaries and the number of strata are different for different BDEs and are determined using some optimization procedure. The analysis variables are salary funds y and number of employees x, taking values $y_1, \dots, y_N$ and $x_1, \dots, x_N$.

The main parameters of SUs change significantly under the changing economic conditions even during one-year period since the complete survey. The main types of changes, which are taken into account are:

a) the enterprise SU changes the type of its economical activity;

b) the number of employees changes (the SU interwieved does not satisfy the requirement for the number of employees which is necessary for the stratum the SU was selected from);

c) the enterprise splits into several new ones;

d) the enterprise joins other units, possibly from different strata.

The practice showed, that 25-30% of SUs of the population of Lithuanian enterprises have changed their parameters during the one year period. For this reason the inclusion probability of the k th unit changes (in some cases, significantly) as compared to the respective probabilities defined at the sampling stage and has to be recalculated. Similarly, the inclusion probabilities π_{kl} are also calculated. Despite that the stratified sampling design was used for sample selection, the general π estimators have to be used in order to get unbiased estimators of totals

$$X = \sum_{k \in D} x_k, \qquad Y = \sum_{k \in D} y_k.$$

The summation is performed over some domain of interest D (union of BDEs) or over the whole population. Thus, the estimators of X and Y are:

$$\widehat{X} = \sum_{k \in S \cap D} \frac{x_k}{\pi_k}, \qquad \widehat{Y} = \sum_{k \in S \cap D} \frac{y_k}{\pi_k}.$$

An attempt is made to use the ratio estimators for the totals X – total number of employees in D and Y – total salary fund in D, exploiting the data of the previous census. Let $\widetilde{x}_k$ and $\widetilde{y}_k$ be the respective values of variables x and y from the previous census. Let us denote

$$\widetilde{X} = \sum_{k \in D} \widetilde{x}_k, \qquad \widetilde{Y} = \sum_{k \in D} \widetilde{y}_k,$$

$$\widehat{\widetilde{X}} = \sum_{k \in S \cap D} \frac{\widetilde{x}_k}{\pi_k}, \qquad \widehat{\widetilde{Y}} = \sum_{k \in S \cap D} \frac{\widetilde{y}_k}{\pi_k}.$$

The ratio estimators of X and Y are

$$\widehat{X}_R = \frac{\widehat{X}}{\widehat{\widetilde{X}}} \widetilde{X}, \qquad \widehat{Y}_R = \frac{\widehat{Y}}{\widehat{\widetilde{Y}}} \widetilde{Y}.$$

These ratio estimators of totals have smaller variances (coefficients of variation) than the corresponding π estimators (see the tables below) because the variables x and $\widetilde{x}$, y and $\widetilde{y}$ are sufficiently correlated.

The main parameter of estimation is the average salary $R = Y/X$ in various domains of estimation (kinds of economic activity). Two estimators of the ratio R have been used:

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}}, \qquad R_R = \frac{\widehat{Y}_R}{\widehat{X}_R}.$$

Using formula (2) of proposition 2 one can get estimators of variances of $\widehat{R}$ and $\widehat{R}_R$:

$$\widehat{\mathbf{D}}\,\widehat{R} = \frac{1}{\widehat{X}^2} \sum_{k,l \in S \cap D} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k - \widehat{R}x_k}{\pi_k} \frac{y_l - \widehat{R}x_l}{\pi_l},$$

$$\widehat{\mathbf{D}}\,\widehat{R}_R = \widehat{R}_R^2 \sum_{k,l \in S \cap D} \frac{\Delta_{kl}}{\pi_{kl}} \frac{\widehat{z}_k}{\pi_k} \frac{\widehat{z}_l}{\pi_l},$$

where

$$z_k = \frac{y_k}{\widehat{Y}} - \frac{x_k}{\widehat{X}} - \frac{\widetilde{y}_k}{\widehat{\widetilde{Y}}} + \frac{\widetilde{x}_k}{\widehat{\widetilde{X}}}.$$

Two mixed estimators

$$\widehat{R}_1 = \frac{\widehat{Y}_R}{\widehat{X}}, \qquad \widehat{R}_2 = \frac{\widehat{Y}}{\widehat{X}_R}.$$

were also examined and their variances were estimated using (2). It was found that estimators $\widehat{R}$ and $\widehat{R}_R$ of the ratio R were preferable, the coefficients of variation of $\widehat{R}_R$ and $\widehat{R}$ were smaller. At the same time the coefficients of variation of the estimators $\widehat{R}$ and $\widehat{R}_R$ were approximately the same.

Some illustrations are presented in the following tables.

The notation

$$cv(\widehat{X}) = \frac{\sqrt{\widehat{\mathbf{D}}\,\widehat{X}}}{\widehat{X}}$$

is used for the coefficient of variation of the estimate $\widehat{X}$.

Table 1.
π estimates and coefficients of variation

Kind of activity	$\widehat{X}$	$\widehat{Y}$	$\widehat{R}$	$cv(\widehat{X})$	$cv(\widehat{Y})$	$cv(\widehat{R})$
Agriculture	60804	19651511	323.2	0.032	0.041	0.024
Construction	48287	36892145	764.01	0.025	0.030	0.020
Transport, storage	44935	36551610	813.43	0.037	0.039	0.014
Services for the community	32194	33163009	1030.11	0.020	0.027	0.017
Financial intermediation	13311	19692176	1479.36	0.022	0.034	0.017
Post and communications	11482	9914808	863.48	0.052	0.053	0.013
Monetary intermediation	10556	16404433	1554.11	0.025	0.035	0.017
Hotels and restaurants	7847	3842186	489.65	0.039	0.039	0.019
Governmental institutions	4110	5022252	1221.90	0.045	0.048	0.021
Fishing	633	271230	428.48	0.042	0.029	0.014

Table 2.
Ratio estimates and coefficients of variation

Kind of activity	$\widehat{X}_R$	$\widehat{Y}_R$	$\widehat{R}_R$	$cv(\widehat{X}_R)$	$cv(\widehat{Y}_R)$	$cv(\widehat{R}_R)$
Agriculture	67732	21101108	311.54	0.020	0.021	0.021
Construction	48362	36850177	761.97	0.023	0.025	0.015
Transport, storage	45125	37473974	830.45	0.012	0.011	0.010
Services for the community	32932	32973581	1001.27	0.016	0.019	0.011
Financial intermediation	13795	20048176	1453.32	0.016	0.025	0.016
Post and communications	12007	10302179	858.00	0.039	0.029	0.012
Monetary intermediation	11067	16935335	1530.31	0.019	0.027	0.016
Hotels and restaurants	8261	4079131	493.79	0.032	0.032	0.018
Governmental institutions	4592	5638992	1227.89	0.015	0.018	0.010
Fishing	642	298490	465.20	0.048	0.049	0.008

REFERENCES

[1] C.-E. Särndal, B. Swenson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, 1992.

Apie regresinius santykio įverčius ir jų taikymą

G. Klimavičius, D. Krapavickaitė, A. Plikusas

Darbe nagrinėjamos specialaus pavidalo statistikos, išreiškiamos sumų įverčių sandaugų santykiu. Tokioms statistikoms rasta apytikrė dispersijos išraiška ir dispersijos statistinis įvertis. Pateikiamas realus tokio pavidalo įverčių taikymo valstybinėje statistikoje pavyzdys.