Liet. matem. rink., **44**, *spec. nr.*, 2004, 547–553

Estimation of variance of complex estimator

Jurgita ČIMŽAITĖ (VGTU), Danutė KRAPAVICKAITĖ (MII)
e-mail: jurgita.cimzaite@mail.std.lt, krapav@ktl.mii.lt

1. Introduction

The aim of this paper is to investigate application of resampling methods for estimation of the variance of nonlinear estimator of the parameter in finite population and for complex sampling design.

Income per capita is one of the parameters which are estimated in household budget survey. Because of stratified two stage sampling design and auxiliary information used at the estimation stage this estimator becomes complex. It is difficult to estimate its variance. It seems that traditionally used method underestimates this variance. The aim of this work is using traditional method, random groups, jackknife and mirror match bootstrap method to estimate this variance and to compare the estimates in order to choose the most suitable method.

The methods applied here can be used in other complex surveys also.

2. Survey population and sampling design

As a survey population $\mathcal{U}$ is used the data of the real household budget survey (HBS) of 2002. It consists of $M = 8\,025$ households. The sampling design, close to the real HBS design in 2002, is used for the simulation study. Population is divided into 7 strata by the place of residence: 5 cities, the rest of the urban area and the rural area. Using the Lithuanian population register as a frame, a simple random sample of individuals is selected in the first 5 strata. The households of the persons selected are included into the sample. Because of unequal number of household members unequal probability sampling design of households is obtained. The household inclusion into the sample probability is

$$\pi_{hj} = \frac{n_h x_{hj}}{M_h},$$

here M_h – number of individuals in the h-th stratum, n_h – sample size of households, x_{hj} – j-th household size, $h = 1, \ldots, 5$.

Two stage cluster sampling design is used in the 6-th and 7-th stratum. Sampling without replacement with household inclusion probabilities proportional to the cluster size is used at the first stage and sampling design like in the first 5 strata is used at the 2-nd stage. The inclusion probability of the j-th household in the i-th cluster from the

h-th stratum equals

$$\pi_{hij} = \pi_{hi}^{(1)} \pi_{hij}^{(2)}, \quad \pi_{hi}^{(1)} = \frac{n_h M_{hi}}{M_h}, \quad \pi_{hij}^{(2)} = \frac{m_{hi} x_{hij}}{M_{hi}},$$

here n_h – sample size of clusters, M_{hi} – number of individuals in the i-th cluster, N_h – number of households in the population for $h = 1, \ldots, 5$ or number of clusters for $h = 6, 7$, $M_h = \sum_{i=1}^{N_h} M_{hi}$ – number of individuals, m_{hi} – household sample size in the i-th cluster, x_{hij} – j-th household size in the i-th cluster of the h-th stratum. Total response is supposed.

3. Parameter and its estimator

Let us denote study variable y – disposable income of the household, x – household size, population totals

$$t_y = \sum_{h=1}^{5} \sum_{j=1}^{N_h} y_{hj} + \sum_{h=6}^{7} \sum_{i=1}^{N_h} \sum_{j=1}^{L_{hi}} y_{hij},$$

$$t_x = \sum_{h=1}^{5} \sum_{j=1}^{N_h} x_{hj} + \sum_{h=6}^{7} \sum_{i=1}^{N_h} \sum_{j=1}^{L_{hi}} x_{hij} = \sum_{h=1}^{7} M_h = M,$$

L_{hi} – number of households in the i-th cluster, M – population size of individuals.

Parameter of interest – income per capita – is expressed as a ratio of two totals:

$$R = \frac{t_y}{t_x}$$

and can be estimated by

$$\widehat{R} = \frac{\widehat{t_y}}{\widehat{t_x}}.$$

Horvitz–Thompson estimator ([2]) can be used for estimation of totals:

$$\widehat{t_y} = \sum_{h=1}^{5} \sum_{j=1}^{n_{hj}} \frac{y_{hj}}{\pi_{hj}} + \sum_{h=6}^{7} \sum_{i=1}^{N_h} \sum_{j=1}^{L_{hi}} \frac{y_{hij}}{\pi_{hij}} = \sum_{k \in s} d_k y_k, \quad \widehat{t_x} = \sum_{k \in s} d_k x_k,$$

s – probability sample, sampling design of which is described in the previous section.

In order the estimates of the number of individuals in some population groups to be equal to the real population constants, calibration of sampling weights is used ([1]). This method allows us to construct the new weights w_k which minimize the distance function

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

and by means of which the known totals t_{z_j} of auxiliary variables z_j, $j = 1, 2, \ldots, J$ can be estimated without error:

$$\sum_{k \in s} w_k z_{jk} = \sum_{k \in \mathcal{U}} z_{jk} = t_{zj}.$$

10 auxiliary variables are used in the study. 7 of them mean the number of individuals in the household belonging to the corresponding stratum. Other indicate the number of individuals in the household belonging to the corresponding age group 5–24, 25–39, 40–60.

Let us denote estimators

$$\widehat{t}_y^{(cal)} = \sum_{k \in s} w_k y_k, \quad \widehat{t}_x^{(cal)} = \sum_{k \in s} w_k x_k, \quad \widehat{R}^{(cal)} = \frac{\widehat{t}_y^{(cal)}}{\widehat{t}_x^{(cal)}}.$$

The estimator of the ratio $\widehat{R}^{(cal)}$ is complex because of sampling design and estimation method.

4. The estimators of the variance of $\widehat{R}^{(cal)}$

4.1. The "true" variance of $\widehat{R}^{(cal)}$

Let us denote by $s_k, k = 1, \ldots, K$ all possible samples under the given sampling design and $p(s_k)$ – the probability the sample s_k to be selected. According to the definition the variance of the estimator $\widehat{R}^{(cal)}$ is a measure of the variability of its values $\widehat{R}_k^{(cal)}$

$$Var\ \widehat{R}^{(cal)} = \mathbf{E}\left(\widehat{R}^{(cal)} - \mathbf{E}\widehat{R}^{(cal)}\right)^2 = \sum_{k=1}^{K} \left(\widehat{R}_k^{(cal)} - \mathbf{E}\widehat{R}^{(cal)}\right)^2 p(s_k).$$

It can not be calculated in practice because of big number K of all possible samples. For this reason its approximation is used. $B = 1000$ probability samples under given sampling design are selected and estimates of the ratio $\widehat{R}_b^{(cal)}$, $b = 1, \ldots, B$ calculated. Their sampling variance

$$V_1 = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{R}_b^{(cal)} - \bar{\widehat{R}}^{(cal)}\right)^2, \quad \bar{\widehat{R}}^{(cal)} = \frac{1}{B} \sum_{b=1}^{B} \widehat{R}_b^{(cal)}$$

is considered as "true" variance of $\widehat{R}^{(cal)}$: $Var\ \widehat{R}^{(cal)} \cong V_1$.

4.2. Traditional variance estimator

Using Taylor linearization method the variance of $\widehat{R}^{(cal)}$ can be expressed by

$$Var\ \widehat{R}^{(cal)} = Var\left(\frac{\widehat{t}_y^{(cal)}}{\widehat{t}_x^{(cal)}}\right) \cong \frac{1}{t_x^2} Var\left(\widehat{t}_y^{(cal)} - R\widehat{t}_x^{(cal)}\right).$$

Let us denote $\widehat{\theta} = \widehat{t}_y^{(cal)} - R\widehat{t}_x^{(cal)}$ – the estimator of the parameter $\theta = t_y - Rt_x$. For two stage sampling design and estimator $\widehat{\theta}$ of parameter θ, the variance

$$Var\ \widehat{\theta} = Var(\mathbf{E}\widehat{\theta}|\mathbf{i}_1) + \mathbf{E}(Var\ \widehat{\theta}|\mathbf{i}_1) \approx Var\ (\mathbf{E}\widehat{\theta}|\mathbf{i}_1)$$

with the 1-st stage sample $\mathbf{i}_1$ can be estimated by

$$\widehat{Var}\ \widehat{\theta} = \widehat{Var}_1\widehat{\theta},$$

here variance Var_1 is the variance of $\widehat{\theta}$ under the 1-st stage sampling design. For this reason the estimator

$$\widehat{Var}\ \widehat{R}^{(cal)} = \frac{1}{t_x^2}\widehat{Var}_1\left(\widehat{t}_y^{(cal)} - R\widehat{t}_x^{(cal)}\right)$$

is used to estimate $Var\ \widehat{R}^{(cal)}$.

Evidently this estimator underestimates the variance. The aim of the simulation study is to get to know how big the underestimation is.

4.3. Random groups method

Using the random groups method ([2]) the initial sample is divided randomly into $A = 4$ groups composed of the elements of all strata. This procedure is applied to the 1-st stage elements, so each of the clusters is included into the group totally.

For each of the group the values $\widehat{R}_a^{(cal)}$ of the parameter R are calculated, $a = 1, \ldots, A$. They are identically distributed, but not independent. Their average is

$$\widehat{R}_{RG}^{(cal)} = \frac{1}{A}\sum_{a=1}^{A}\widehat{R}_a^{(cal)}$$

with the variance

$$Var\ \widehat{R}_{RG}^{(cal)} \cong \frac{1}{A^2}\sum_{a=1}^{A} Var\ \widehat{R}_a^{(cal)} = \frac{\sigma^2}{A},$$

here $\sigma = Var\ \widehat{R}_a^{(cal)}$. Estimator

$$\widehat{Var}\ \widehat{R}_{RG}^{(cal)} = \frac{1}{A(A-1)}\sum_{a=1}^{A}\left(\widehat{R}_a^{(cal)} - \widehat{R}_{RG}^{(cal)}\right)^2$$

Table 1. Composition of the random groups

Group:	1	2	3	4	5	6	7
Number of the 1-st stage elements							
in the sample:	80	56	24	20	16	8	24
in the group:	20	14	6	5	4	2	6

is used to estimate $Var\,\widehat{\theta}$. This estimator is biased.

4.4. Jackknife method

There are $A = 228$ first stage sampling elements in the sample. For any $a = 1, \ldots, A$, the a-th first stage sampling unit is deleted from the sample, and estimator $\widehat{R}_{(a)}^{(cal)}$ of R is calculated using the rest of the sample. Afterwards the pseudo-values

$$\widehat{R}_a^{(cal)} = n\widehat{R}^{(cal)} - (n-1)\widehat{R}_{(a)}^{(cal)},$$

$a = 1, \ldots, A$ are calculated and alternative estimator for R is build:

$$\widehat{R}_{jack}^{(cal)} = \frac{1}{A}\sum_{a=1}^{A} \widehat{R}_a^{(cal)}.$$

Estimator of its variance

$$\widehat{Var}\,\widehat{R}_{jack}^{(cal)} = \sum_{h=1}^{7}\left(1 - \frac{n_h}{N_h}\right)\frac{n_h - 1}{n_h}\sum_{i=1}^{n_h}\left(\widehat{R}_{(hi)}^{(cal)} - \widehat{R}^{(cal)}\right)^2$$

is used to estimate the variance $Var\,\widehat{R}^{(cal)}$: $\widehat{Var}_{jack}\,\widehat{R}^{(cal)} = \widehat{Var}\,\widehat{R}_{jack}^{(cal)}$ and called the jackknife estimator ([2]). Here $\widehat{R}_{(hi)}^{(cal)}$ means the estimator of R obtained when the i-th first stage sampling unit from the h-th stratum is deleted.

4.5. Mirror match bootstrap

Mirror match bootstrap ([3], [4]) is applied in subsequent steps:

1. The simple random subsample of the first stage sampling units of size

$$n_h^\star = \left[\frac{n_h}{2 - f_h}\right] < n_h,$$

$f_h = n_h/N_h$ is drawn from the initial sample of each stratum.

2. The elements of the subsample are replaced in the sample and the 1-st step is repeated random number K_h of times:

$$P(K_h = k_{h1}) = p_h, \quad P(K_h = k_{h2}) = 1 - p_h,$$

$$k_{h1} = \left[\frac{n_h(1 - f_h^\star)}{n_h^\star(1 - f_h)}\right], \quad k_{h2} = k_{h1} + 1, \quad p_h = \frac{\frac{1}{k_h} - \frac{1}{k_{h1}}}{\frac{1}{k_{h1}} - \frac{1}{k_{h2}}}, \quad f_h^\star = n_h^\star/n_h,$$

$h = 1, 2, \ldots, 7$. Thus, the new sample in each of the stratum of size $m_h = K_h n_h^\star$ is obtained, which is called the bootstrap sample. Let us denote the values of study variables of the bootstrap sample

$$y_{h1}^\star, \ldots, y_{hm_h}^\star, \quad x_{h1}^\star, \ldots, x_{hm_h}^\star.$$

Table 2. Bootstrap sample sizes and sampling fractions

Stratum	Number of households	N_h	n_h	f_h	$n_h^\star$	$f_h^\star$	K_h	$K_h n_h^\star$
1	1 163	1 163	80	0.07	27	0.34	2	54
2	781	781	56	0.07	19	0.34	2	38
3	338	338	24	0.07	8	0.33	2	16
4	295	295	20	0.07	6	0.30	3	18
5	250	250	16	0.06	5	0.31	2	10
6	2091	20	8	0.40	4	0.50	3	12
7	3 107	66	24	0.36	8	0.33	3	24

Estimates of the population totals $\widehat{t}_y^{(cal)\star}$, $\widehat{t}_x^{(cal)\star}$ are calculated using this data and population ratio R is estimated by bootstrap estimator $\widehat{R}^{(cal)\star} = \widehat{t}_y^{(cal)\star} / \widehat{t}_x^{(cal)\star}$.

3. Steps **1** and **2** are repeated $B = 100$ times, estimates of the ratio $\widehat{R}_1^{(cal)\star}, \ldots, \widehat{R}_B^{(cal)\star}$ are obtained. The variance $Var\,\widehat{R}^{(cal)}$ of the estimator $\widehat{R}^{(cal)}$ is estimated by the variance of the estimator $\widehat{R}^{(cal)\star}$ of R

$$Var_\star \widehat{R}^{(cal)\star} = E_\star\left(\widehat{R}^{(cal)\star} - \mathbf{E}_\star \widehat{R}^{(cal)\star}\right)^2$$

under the bootstrap sampling design or by its mirror match bootstrap estimator

$$\widehat{Var}_{MM}\widehat{R}^{(cal)} = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{R}_b^{(cal)\star} - \bar{\widehat{R}}^{(cal)\star}\right), \quad \bar{\widehat{R}}^{(cal)\star} = \frac{1}{B}\sum_{b=1}^{B}\widehat{R}_b^{(cal)\star}.$$

The choice of sample sizes and sampling fractions of the simulation study is presented in Table 2.

Simple random subsampling in strata is used in the simulation despite the original sampling design in the strata is with unequal probabilities.

5. Simulation results

There were drawn 100 samples from the population under the given sampling design. The variance estimates of the estimator of the parameter of interest where calculated in each case.

Let us define a relative bias of the estimator of the variance of the estimator $\widehat{\theta} = \widehat{R}^{(cal)}$

$$Relative\ \ bias\,(\widehat{Var}\,\widehat{\theta}) = \frac{\widehat{Var}\,\widehat{\theta} - Var\,\widehat{\theta}}{Var\,\widehat{\theta}}$$

$$Relative\ \ MSE\,(\widehat{Var}\,\theta) = \frac{\sqrt{(\widehat{Var}\,\widehat{\theta} - Var\,\widehat{\theta})^2 + \widehat{Var}\,(\widehat{Var}\,\widehat{\theta})}}{Var\,\widehat{\theta}}.$$

Summary of the estimates obtained is presented in the Table 3.

Variability of the estimator due to the first stage sampling design is investigated.

Table 3. Estimates of the variances $Var\ \widehat{R}^{(cal)}$ of the estimators

Method	Average	Min	Max	Variance	Relative bias	Relative MSE
"True" variance	160.51					
Traditional	112.48	49.60	268.69	1944.52	−0.30	0.41
Random groups	188.68	1.82	1299.47	34197.96	0.18	1.17
Jackknife	144.58	85.69	283.47	954.83	−0.10	0.21
Bootstrap	218.30	119.66	297.94	2312.31	0.36	0.46

Following conclusions can be made:

1. The "true" variance of the estimator of income per capita is underestimated using traditional estimator.
2. The variance of the estimator of variance obtained with the random groups method is unacceptable big. The small number of groups (only 4) can be reason for this.
3. The MSE of the bootstrap estimator is quite big.
4. The estimator of variance obtained using jackknife method has the smallest MSE and seems to be the mostly acceptable.

Acknowledgments. The authors are thankful to the Statistics Lithuania for the possibility to analyse the survey data.

References

1. J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382 (1992).
2. C.-E. Särndal, B. Swensson, J. Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, New York (1992).
3. J. Shao, D. Tu, *The Jackknife and Bootstrap*, Springer-Verlag, New York (1995).
4. R.R. Sitter, Comparing three bootstrap methods for survey data, *The Canadian Journal of Statistics*, **20**(2), 135–154 (1992).

REZIUMĖ

J. Čimžaitė, D. Krapavickaitė. Sudėtingo įvertinio dispersijos vertinimas

Šio darbo tikslas – ištirti kartotinių imčių metodų taikymo galimybes netiesinio sumų atžvilgiu paramero įvertinio dispersijai vertinti, esant sudėtingam imties planui.

Naudojant realius oficialiosios statistikos namų ūkių tyrimo duomenis, vertinama vieno namų ūkio nario vidutinių pajamų įverčio dispersija, naudojant tradicinį, atsitiktinių grupių, džeknaifo ir butstrepo metodus. Darbas padės parinkti tyrimui tinkamiausią dispersijos vertinimo būdą. Nagrinėjami metodai gali būti taikomi ir kituose sudėtinguose imčių tyrimuose.