Lithuanian Journal of Statistics *2012, vol. 51, No 1, pp. 17-21*
Lietuvos statistikos darbai *2012, 51 t., Nr. 1, 17–21 p.*
www.statisticsjournal.lt

CONFIDENCE INTERVALS IN OFFICIAL STATISTICS: THE CASE OF LITHUANIA

Andrius Čiginas

Statistics Lithuania. Address: 29 Gedimino Ave., Vilnius, Lithuania. E-mail: andrius.ciginas@stat.gov.lt
Vilnius University, Institute of Mathematics and Informatics. Address: 4 Akademijos St., Vilnius, Lithuania

Received: May 2012 Revised: September 2012 Published: November 2012

Abstract. In the paper, the application of confidence intervals in the surveys of official statistics is discussed. It is noticed that there are situations where at the first sight natural normal distribution-based confidence intervals are not suitable. We demonstrate it by examples taken from Lithuanian statistical surveys. We also discuss an Edgeworth expansion and a bootstrap method as an alternative to the normal approximation.

Keywords: confidence interval, normal approximation, Edgeworth expansion, bootstrap, finite population.

1. Introduction

A very common parameter estimated, for example, in Lithuanian surveys of official statistics is the sum of measurements of a certain variable of interest in a finite population of enterprises or individuals. A typical sample design used to form a sample is stratified simple random sampling, where, e.g. in cases of surveys of enterprises, the strata are naturally formed by economic activity and by sizes of enterprises. The next thing which is always important in the estimation process and which must be controlled is the quality of estimates. There are several very common ways to present it. The first way is the estimate of the coefficient of variation (or variance), the second one, which is somewhat more informative, is the estimate of the confidence interval. In this paper, we discuss the use of the latter one.

The application of the traditional normal confidence interval

$$(\hat{\theta} + z_{\alpha/2}\sqrt{Var(\hat{\theta})}, \hat{\theta} + z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}) \tag{1}$$

is based on the assumption that the estimator $\hat{\theta}$ has a normal or approximately normal distribution. Here $Var(\hat{\theta})$ is the variance of $\hat{\theta}$ and z_α denotes α-quantile of the standard normal distribution. In case of large-scale surveys (where sample sizes are large), the normality assumption is quite natural because, theoretically, many estimators (i.e. not only estimators of sums) are asymptotically normal under quite mild assumptions on the population. In particular, for simple random samples without replacement (one-stratum case) the central limit theorem, in the case of the sample mean, was proved in [5]. For the case of classical linear combinations of stratum means (sums) in stratified sampling:

$$\hat{\theta} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_{hj}\,, \tag{2}$$

we refer to e.g. [2]. Here N_h and n_h are the population and sample sizes in h-stratum; y_{hj}, $j = 1,\ldots,n_h$ denote the sample values of the study variable y in h-stratum.

2. Numerical examples

Example 1. We illustrate the use of (1) in the following numerical example. Let the population of interest be two strata of medium-size enterprises from the survey of construction. The sizes of the strata are $N_1 = 731$ and $N_2 = 455$. The corresponding sample sizes are $n_1 = 136$ and $n_2 = 204$. We aim to estimate the population sum of the number of employees (here we use the data for the fourth quarter of 2011). Since the values of this study variable are known for

all units of the population (from administrative data sources), we evaluate the distribution of the estimator (2) and compare it with the normal distribution. Specifically, we estimate the distribution

$$F(x) = P\{\hat{\theta} - E(\hat{\theta}) \le x\sqrt{Var(\hat{\theta})}\} \tag{3}$$

by applying the Monte Carlo method, i.e. by drawing independently C stratified samples from the population and creating the empirical distribution from the standardized observations $(\hat{\theta}_i - \hat{\mu})/\hat{s}$, where $\hat{\mu}$ and $\hat{s}^2$ are the empirical mean and variance of $\hat{\theta}_i$, $i = 1,\ldots,C$, respectively. We take $C = 1000$. It is seen from Fig. 1, and the normality tests show that, in this case, the distribution (3) is close to the standard normal. Thus, the use of z_α in (1) has a background.

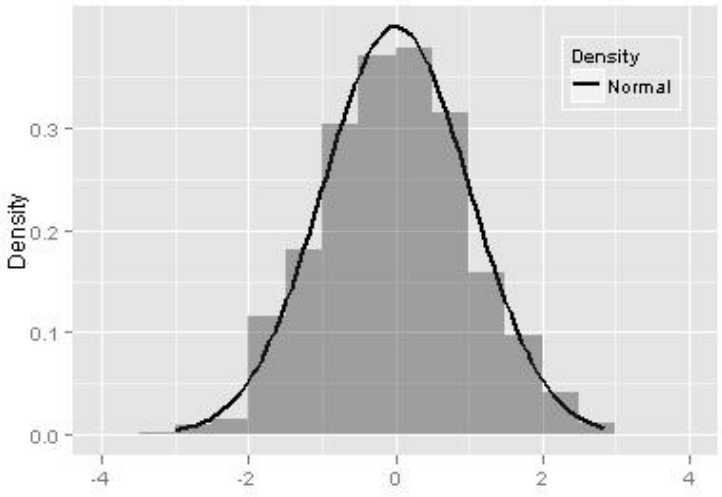

Fig. 1. Distribution of the estimator in the construction survey

Example 2. To show a different situation, the data are taken from the survey on investment. The parameter of interest is the population sum of investment in tangible fixed assets. We form an artificial population from the sample data of several strata (data for the fourth quarter of 2011). The size of the whole new population is $N = 665$, and the total sample size is chosen to be $n = 200$. We evaluate the distribution (3) of the estimator (2) in the same way as in the previous example, and now results are different: the normality tests and Fig. 2 show that the distribution is not close to the standard normal. Next, the evaluated quantiles $q_{\alpha/2}$ and $q_{1-\alpha/2}$ of the distribution (3), if $\alpha = 0.05$, are $q_{0.025} = -1.52$ and $q_{0.975} = 2.08$.

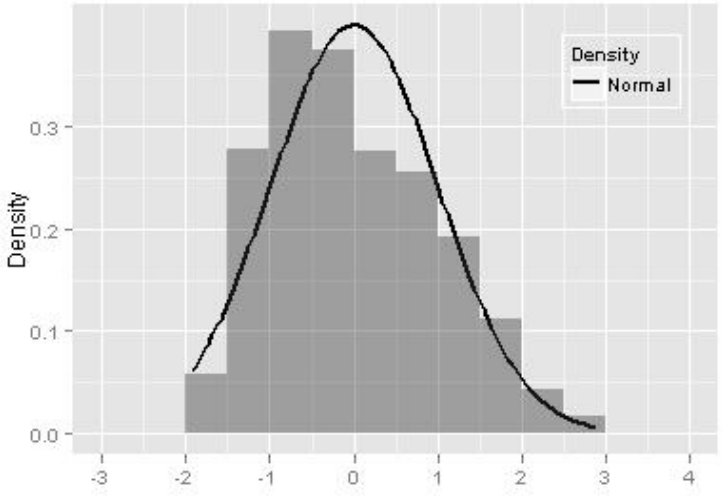

Fig. 2. Distribution of the estimator in the investment survey

The difference between these two examples of sample surveys lies in the different distributions of the populations. It is seen from Fig. 3 and Fig. 4 that the population values of investment in tangible fixed assets are much more asymmetric compared to the values of the first population. Recall (e.g. from [5]) a Lindeberg-type Erdős–Rényi condition, which also means that the survey population should not be very asymmetric.

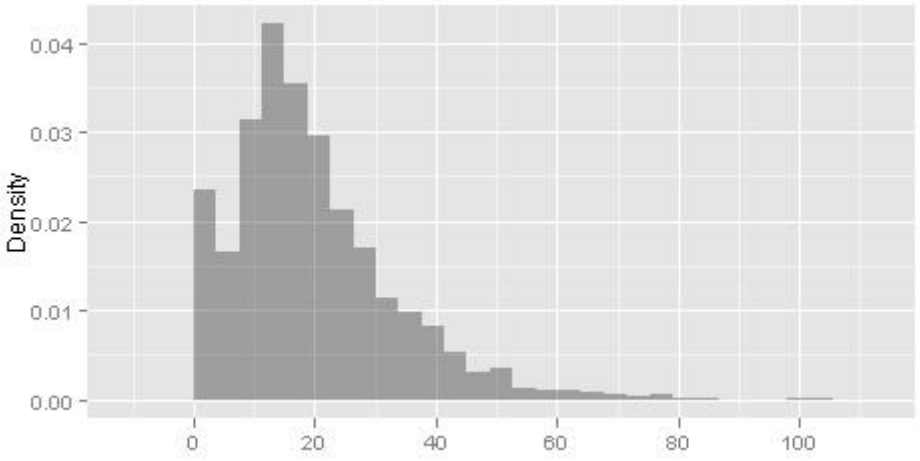

Fig. 3. Number of employees in the population of the construction survey

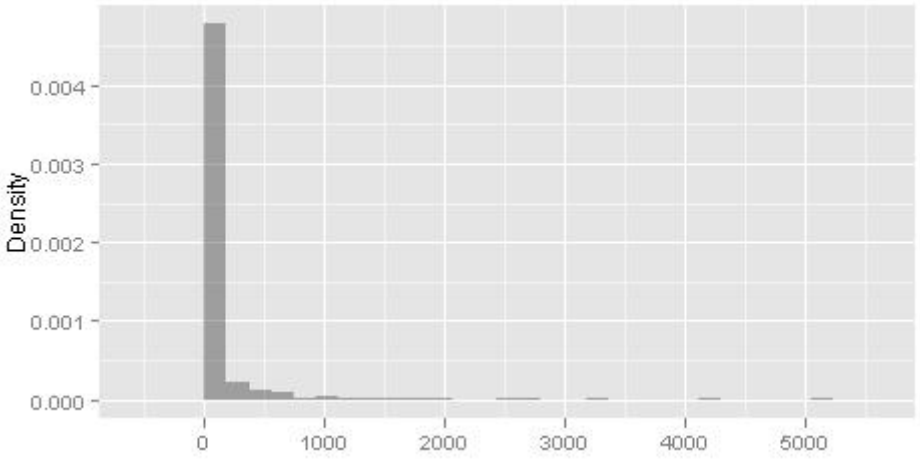

Fig. 4. Investment in tangible fixed assets in the population of the investment survey

Thus, in the case of data of Example 2, we need to use methods other than the normal approximation to estimate (3).

3. Edgeworth and bootstrap approximations

The two well-known second-order approximations to (3) are Edgeworth expansions and bootstrap approximations. In particular, the one-term (short) Edgeworth expansion is of the form

$$G(x) = \Phi(x) - \frac{\lambda}{6}(x^2 - 1)\phi(x), \tag{4}$$

where $\Phi(x)$ and $\phi(x)$ denote the distribution and density of the standard normal random variable, and λ is, in fact, (an approximation to) the standardized third cumulant of the estimator of interest. The Edgeworth correction term, added to the normal approximation in (4), reflects the skewness of the distribution of the estimator and thus an asymmetry of the population. In the case of the estimator (2), the one-term Edgeworth expansion was studied in [1]; see

also [4]. It is important to note that the theory in [1] (and [4]) also holds for other common estimators: ratio and regression estimators in stratified samples. It is well known that typically $|F(x)-\Phi(x)|=O(n^{-1/2})$, while in the case of the short Edgeworth expansion (4), in many situations,

$$|F(x)-G(x)|=o(n^{-1/2}) \tag{5}$$

holds. Clearly, the parameter λ is an unknown characteristic of the population and should be estimated. Then, with the consistent estimate $\hat{\lambda}$ of λ, (5) holds in probability.

The second universal method is the bootstrap. There are several bootstrap variants considered in literature for the case of samples without replacement. Some of these methods are reviewed in [4]. In the present paper, we discuss one method, which is proposed in [4]. The practical realization of this method is the following. For each $h=1,\ldots,H$, write $N_h=m_h n_h+t_h$, $0\le t_h<n_h$. Then, for each $h=1,\ldots,H$, we form an empirical stratum by repeating m_h times the sample $\mathrm{Y}_h=\{y_{hj}, j=1,\ldots,n_h\}$ and joining the simple random sample without replacement z_{hj}, $j=1,\ldots,t_h$ drawn from Y_h. Then the union of the empirical strata is one empirical population. Next, in the same way as in the evaluation of the distribution (3) by Monte Carlo method, we draw R samples (called resamples) from the empirical population and calculate standardized values of the estimator of interest. Denote them by $\widetilde{\theta}_i^{(1)}$, $i=1,\ldots,R$. Next, we repeat it with other $B-1$ empirical populations and, finally, from the collection of data $\widetilde{\theta}_i^{(b)}$, $i=1,\ldots,R$, $b=1,\ldots,B$, we get the (empirical) bootstrap estimate $H(x)$ of (3). It is shown in [4] that, for many common estimators of the sum (including (2)), under certain conditions, $H(x)$ is the second-order correct approximation, i.e. $|F(x)-H(x)|=o_P(n^{-1/2})$. Moreover, it is well known that the bootstrap approximation $H(x)$ is in a sense very similar to the one-term Edgeworth expansion $G(x)$, since it also captures the skewness of the distribution of the estimator.

Example 2 (continued). We apply the bootstrap approximation $H(x)$ to $F(x)$. To see how it works, we draw 80 stratified samples from the population and, for each of them, we calculate $H(x)$ by the algorithm given above (we take $R=100$ and $B=10$). Then we get 80 pairs of 0.025 and 0.975 quantiles of $H(x)$ and, in order to see the efficiency of the bootstrap method, we present histograms for both quantiles (see Fig. 5).

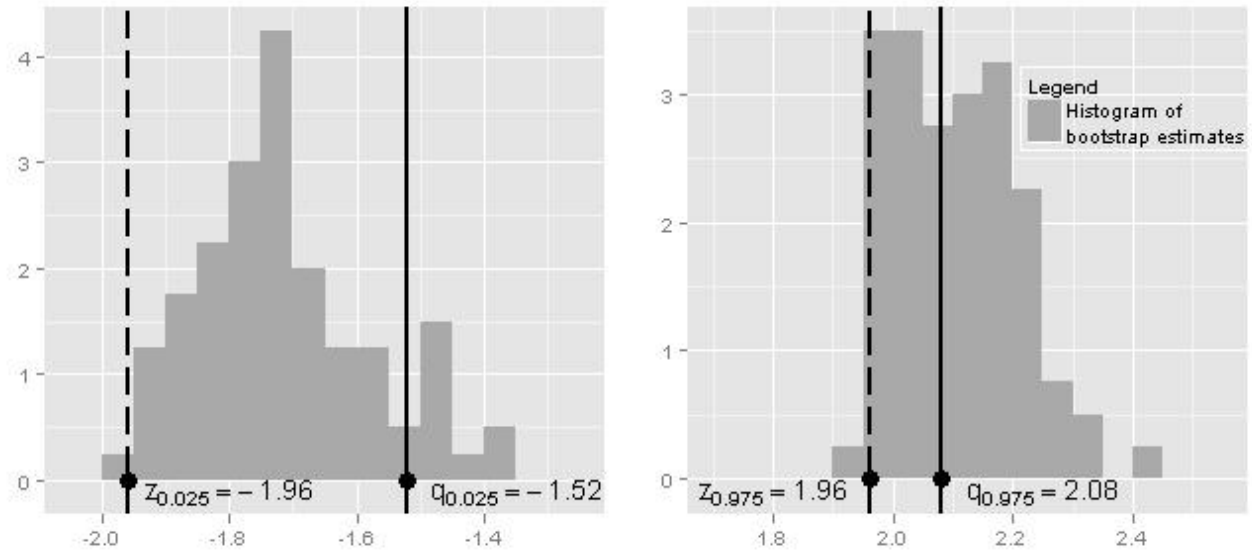

Fig. 5. Normal and the evaluated $F(x)$ quantiles and histograms of bootstrap quantiles

It is seen that, applying a bootstrap approximation, there is a very small risk to get worse estimates of quantiles than the normal quantiles. Note that, by the obtained results, the corresponding bootstrap confidence interval

$(\hat{\theta}+\hat{q}_{\alpha/2}\sqrt{Var(\hat{\theta})},\hat{\theta}+\hat{q}_{1-\alpha/2}\sqrt{Var(\hat{\theta})})$, where $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ are the bootstrap estimates of $q_{\alpha/2}$ and $q_{1-\alpha/2}$, tends to be conservative.

4. Remarks

Note that, in the real applications of the confidence interval $(\hat{\theta}+q_{\alpha/2}\sqrt{Var(\hat{\theta})},\hat{\theta}+q_{1-\alpha/2}\sqrt{Var(\hat{\theta})})$, we need to estimate not only $q_{\alpha/2}$ and $q_{1-\alpha/2}$, replacing them by the normal quantiles as in (1) or by the quantiles from the Edgeworth expansion or bootstrap distribution, but we also need to estimate the variance $Var(\hat{\theta})$. Here we do not consider either this question or the estimation of the parameter λ, which defines the short Edgeworth expansion (4).

It is important to mention that there are some works on second-order approximations, which are done for more general (than traditional estimators of sums) estimators in the case of stratified samples, see e.g. [3] on U-statistics (which are close to very general symmetric statistics), but a number of important results are still not extended from the one-stratum case to the stratified sampling or more complex sampling designs. On the other hand, sometimes we need to estimate the accuracy of the estimate in a single stratum, where the sample size is small and therefore the normal approximation often fails.

References

1. Babu G. J. and Singh K. Edgeworth expansions for sampling without replacement from finite populations, *J. Multivariate Anal.*, 17, p. 261–278, 1985.
2. Bickel P. J. and Freedman D. A. Asymptotic normality and the bootstrap in stratified sampling, *Ann. Statist.*, 12, p. 470–482, 1984.
3. Bloznelis M. Second-order and resampling approximation of finite population U-statistics based on stratified samples, *Statistics*, 41, p. 321–332, 2007.
4. Booth J., Butler R. and Hall P. Bootstrap methods for finite populations, *J. Amer. Statist. Assoc.*, 89, p. 1282–1289, 1994.
5. Erdős P. and Rényi A. On the central limit theorem for samples from a finite population, *Publ. Math. Inst. Hungar. Acad. Sci.*, 4, p. 49–61, 1959.

PASIKLIAUTINIEJI INTERVALAI OFICIALIOJOJE STATISTIKOJE: LIETUVOS ATVEJIS

Andrius Čiginas

Santrauka. Straipsnyje aptariamas pasikliautinųjų intervalų vertinimas oficialiosios statistikos tyrimuose. Atkreipiamas dėmesys į atvejus, kada įprasti normaliuoju skirstiniu pagrįsti pasikliautinųjų intervalų įvertiniai yra nepakankamai tikslūs. Tai iliustruojama Lietuvos statistinių tyrimų pavyzdžiais. Siūlomi alternatyvūs Edgewortho skleidinio ir savirankos metodais pagrįsti pasikliautinųjų intervalų įvertiniai.

Reikšminiai žodžiai: pasikliautinasis intervalas, normalioji aproksimacija, Edgewortho skleidinys, saviranka, baigtinė populiacija.