ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Dorota Pekasiewicz**

APPLICATION OF SIMULATION METHODS TO ESTIMATION OF VARIANCE OF NONPARAMETRIC SEQUENTIAL ESTIMATOR OF MEAN

Abstract

Nonparametric sequential methods allow to estimate unknown parameter of random variable distribution, when the distribution of the variable is unknown. We can apply these methods to different sampling designs.

This paper contains a proposal of applying simulation methods to estimate the variance of a nonparametric estimator of mean. An application of bootstrap methods to estimate the variance of a synthetic estimator of the mean in sequential estimation is also presented.

Key words: sequential estimation, bootstrap method, synthetic estimator.

I. INTRODUCTION

Nonparametric sequential methods allow to estimate unknown probability distribution function of random variable under investigation, as well as an unknown distribution parameter, when the distribution class is unknown.

To estimate distribution parameters e.g. the mean, sequential procedures of nonparametric point or interval estimation can be applied for different sequential sampling schemes.

The idea of sequential point estimation of the mean is to determine its estimator values from random sample with size that minimizes risk function. If we do not take into account sampling costs connected with the sequential drawing of elements, the risk function is equal to the mean square error and, in the case of unbiased estimators, to relevant estimator's variance. In such cases we determine the estimation precision through establishing

* Ph.D., Chair of Statistical Methods, University of Łódź.

the number which cannot be exceeded by variance or the mean square error. Sample size is being increased sequentially until the desired precision of estimation is achieved.

Calculating the variance of parameter estimator at every stage of the sequential procedure is not always easy, sometimes even impossible. Due to the complicated form of some parameter estimators often, we do not have any information about their variance or about variance estimators. In the application of the estimators of this kind to sequential estimation of the mean there appears a problem of defining the stopping procedure for the sample size increasing process. In this paper we suggest using, in such cases, simulation methods of estimating variance such as Mahalanobis method, jackknife or bootstraps. An example of applying bootstrap method to estimate the variance of synthetic estimator of subpopulation mean is also presented.

II. NONPARAMETRIC SEQUENTIAL POINT ESTIMATION

Let X be random variable and θ be unknown mean value of this variable. By $\hat{\theta}_n$ we denote an estimator of parameter θ determined from sample $X_1, ..., X_n$.

At every stage of a sequential process aiming at assessing the value of θ we face the statistical problem of decision making. We decide about increasing the sample by drawing one or a few more elements or about concluding the sampling process and treating the estimator's value that we arrived at, as a good enough estimate of θ.

When we make a decision we can define the loss function as follows (see Sen, 1984):

$$L_n = g(|\hat{\theta}_n - \theta|) + c(n), \tag{1}$$

where g is a nonnegative, nonincreasing function on $(0, +\infty)$, with property $g(0) = 0$, and $c(n)$ is the cost function associated with drawing sample.

Let us assume that $n_0 (n_0 \geqslant 1)$ exists, such that for every $n \geqslant n_0$ $\mathrm{E}[g(|\hat{\theta}_n - \theta|)]$ exists.

The risk incurred in the estimation of the mean θ from an n-element sample is given by the formula:

$$R_n = \mathrm{E}L_n = \mathrm{E}[g(|\hat{\theta}_n - \theta|)] + c(n) \quad \text{for} \quad n \geqslant n_0. \tag{2}$$

The quantity R_n may be viewed as the sum of two functions of argument n. The function c is nondecreasing (e.g. $c(n) = pn$, where p denotes the cost of drawing one element), and if $\hat{\theta}_n$ is a consistent estimator of θ, the function $E[g(|\hat{\theta}_n - \theta|)]$ is nonincreasing, monotone and converges to 0 (see Sen, 1984).

We make a decision at such a sample size n for which the function R_n reaches its minimum. For established functions g and c we define n^* as follows:

$$n^* = \min\{n \geqslant n_0: \quad R_n = \inf_m R_m\}. \tag{3}$$

The value of estimator $\hat{\theta}_{n^*}$ is the minimum risk estimate of parameter θ.

If, in the process of the sequential point estimation of θ, we do not take into account the sampling cost, the risk incurred in the sequential estimation of the parameter from n-element sample will depend only on $E[g(|\hat{\theta}_n - \theta|)]$.

If $g(x) = x^2$, it implies $R_n = E(|\hat{\theta}_n - \theta|^2)$ and for an unbiased estimator $\hat{\theta}_n$, function R_n will be the variance of parameter θ estimator ($R_n = D^2(\hat{\theta}_n)$), while for a biased estimator $\hat{\theta}_n$, function R_n will be equal to the mean square error ($R_n = D^2(\hat{\theta}_n) + E(|E(\hat{\theta}_n) - \theta|^2)$).

The sequential point estimation of parameter θ with the aid of an unbiased estimator will be characterized with the stopping rule for the drawing process determined by inequality $D^2(\hat{\theta}_n) \leqslant \varepsilon^2$, where ε is a fixed estimation precision. That means that we will be sequentially adding elements to the sample as long as estimator's variance is less or equal ε^2. In such cases size n^* ensuring the estimation of parameter θ with precision not smaller than the fixed one, is defined in the following way:

$$n^* = \min\{n: \quad \hat{D}^2(\hat{\theta}_n) \leqslant \varepsilon^2\}, \tag{4}$$

where $\hat{D}^2(\hat{\theta}_n)$ is an estimator of variance $D^2(\hat{\theta}_n)$. The value of estimator $\hat{\theta}_{n^*}$ is the estimate of parameter θ with precision not exceeding ε.

III. SIMULATION METHODS OF VARIANCE ESTIMATION

We use simulation methods in estimating the variance of estimators of the mean, when we do not know neither the variance nor any variance estimator (usually due to the complicated formula of the estimator of the mean). The use of such estimators in sequential estimation is possible, but

we encounter the problem of defining the stopping rule. If we do not take into account the sampling costs, the sequential estimation procedure will be connected with estimating variance (mean square error), at every stage, and comparing it with a fixed estimation precision.

To assess the variance of the estimators of the mean the following simulation methods for every stage of the sequential procedure are proposed:

- the Mahalanobis method;
- the jackknife method;
- the bootstrap method.

Mahalanobis method. In the first step of the Mahalanobis procedure in the sequential estimation of parameter θ from the drawn k_1-element sample we create s disjoint subsamples ($s \geqslant 2$), containing l_1^i elements, for $i = 1, 2, ..., s$. If sample size k_n ($n = 1, 2, ...$) can be devided by a fixed number s, the subsample sizes are determined from the formula:

$$l_n^i = \frac{k_n}{s} \quad \text{for} \quad i = 1, ..., s. \tag{5}$$

In other case the sizes of particular subsamples are given by the formula:

$$l_n^i = \begin{cases} \left[\frac{k_n}{s}\right] + 1 & \text{for} \quad i = 1, ..., c \\ \left[\frac{k_n}{s}\right] & \text{for} \quad i = c+1, ..., s \end{cases} \tag{6}$$

where $c = k_n - s\left[\frac{k_n}{s}\right]$.

From each of the s subsamples we determine the value of estimator $\hat{\theta}_{l_1^i}$ ($i = 1, ..., s$), then from all samples containing k_1 elements we calculate the value of estimator $\hat{\theta}_n$ for $n = 1$. The variance of this estimator is assessed with the formula:

$$\hat{\mathrm{D}}^2(\hat{\theta}_n) = \frac{1 - \frac{k_n}{N}}{s(s-1)} \sum_{i=1}^{s} (\hat{\theta}_{l_n^i} - \hat{\theta}_n)^2 \quad \text{for} \quad n = 1, 2, ..., \tag{7}$$

where N denotes the population size.

If the assessed variance value is less or equal to a fixed number ε^2, the value of estimator $\hat{\theta}_1$, determined from k_1 elements, constitutes a good estimate of the mean of the random variable considered. In other case we enlarge the sample by d elements ($d = 1, 2, ...$). At the n-th stage sample

will have $k_n = k_1 + (n-1)d$ elements. The subsample sizes are determined from formula (5) or (6). The value of estimator $\hat{\theta}_n$ is determined from k_n-element sample and its variance from s values of the estimator determined from l_n^i-element subsamples. The sequential point estimation procedure is repeated until the value of the variance estimator of the estimator used is less or equal to ε^2.

Jackknife method. In the first step ($n = 1$) we draw according to an arbitrary scheme, but not the layer one, k_1 population elements and similarly as in the Mahalanobis method we create s subsamples. However, these subsamples are created in different way. We randomly remove from the k_1-element sample l_1^i elements from formula (5) or (6), respective to k_1 being divisible by s or not.

The variance of the mean estimator is estimated from the subsamples consisting of $k_n - l_n^i$ elements on the basis of the formula (see Bracha, 1998):

$$\hat{\mathrm{D}}^2(\hat{\theta}_n) = \frac{1}{s(s-1)} \sum_{i=1}^{s} (\hat{\theta}'_{l_n^i} - \hat{\theta}_n)^2, \tag{8}$$

where

$$\hat{\theta}'_{l_n^i} = s\hat{\theta}_n - (s-1)\hat{\theta}'_{l_n^i}, \tag{9}$$

where $\hat{\theta}_n$ and $\hat{\theta}_{l_n^i}$ are the estimators of the mean determined from the k_n-element proper sample and $(k_n - l_n^i)$-element i-th subsample ($i = 1, ..., s$), respectively.

If $\hat{\mathrm{D}}^2(\hat{\theta}_n)$ is greater then ε^2, we enlarge the sample by d elements and repeat the above procedure.

When we apply the Mahalanobis or jackknife method to assess the variance of the estimator considered we are encountered with the problem of the number s of subsamples determined at the begining and the following steps of sequential estimation. If the sample size at a certain stage of the sequential procedure grows considerably with respect to the initial sample then s should be changed. In such cases the use of the Mahalanobis and jackknife method is more problematic.

Another method that may be useful to assess the variance of the estimator in sequential estimation is the bootstrap method.

Bootstrap method. In the first step of the sequential estimation of sample we draw k_1 elements from the population. These sample observation allow to determine the value of estimator $\hat{\theta}_n$ for $n = 1$. Then from the existing sample we generate J (e.g. $J = 1000$) realizations of the bootstrap sample i.e. the sample generated according to the bootstrap distribution:

$$P(X_b = x_m) = \frac{1}{k_n} \quad \text{for} \quad m = 1, ..., k_n \quad \text{and} \quad b = 1, 2, ..., J \tag{10}$$

We determine the value of the estimator $\hat{\theta}_{b,n}$ for $n = 1$ and $b = 1, 2, ..., J$. The estimator's variance is assessed with the formula:

$$\hat{D}^2(\hat{\theta}_n) = \frac{1}{J} \sum_{b=1}^{J} (\hat{\theta}_{b,n} - \hat{\theta}_n)^2 \tag{11}$$

(in the first stage we assume $n = 1$).

If the condition $\hat{D}^2(\hat{\theta}_n) \leqslant \varepsilon^2$ does not hold, we draw the fixed number d of elements, pool them and the sample together, arriving at the sample consisting of $k_{n+1} = k_n + d$ elements, for $n = 1, 2, 3, ...$ For the pooled sample we determine J realizations of the bootstrap sample and we assess the estimator's variance. We go on with the described process until the variance assessment does not exceed ε^2.

IV. NONPARAMETRIC SEQUENTIAL ESTIMATION OF SUBPOPULATION MEAN

Let us consider the problem of the estimation of the mean in some distinguished subpopulations of the whole population, when we do not know its distribution. If we have some information about the values of the random variable in the whole population as well as about an auxiliary variable correlated with the variable considered, we may use it in synthetic estimators, which are more effective than direct estimators i.e. determined from the subpopulation sample (see Dol, 1991).

Synthetic estimators are constructed on assumption that the parameters of the distribution of the variable investigated in subpopulation are very close to the parameters of the distribution of this variable in the whole population.

Let us denote the variable investigated by X and the auxiliary variable by Y. Moreover, let us assume that the population and subpopulation are divided into G layers.

One of synthetic estimators of the mean θ_0 of variable X for subpopulation is given by the formula:

$$\hat{\theta}_0^n = \frac{1}{N_0} \sum_{g=1}^{G} \frac{\hat{\bar{X}}_g^n}{\bar{Y}_g} TY_{0g}, \tag{12}$$

where N_0 is the subpopulation size, TY_{0g} – the global value of the auxiliary variable Y in the g-th layer of the subpopulation, $\overline{Y}_g$ – the mean value of variable Y in the g-th layer of the population, $\hat{\overline{X}}_g^n$ – the mean value of variable X in the g-th layer of the population estimated from k_n-element sample of the whole population.

We start estimation from the k_1-element sample. We calculate the value of the estimator given by formula (12) and by means of the bootstrap method we estimate its variance. If the variance does not exceed ε^2 we conclude the estimation procedure judging the value of estimator (12) we got as a good enough estimate of the subpopulation mean of variable X. Otherwise, we draw new elements and we repeat the whole procedure.

V. EXAMPLE OF THE APPLICATION OF BOOTSTRAP METHOD TO VARIANCE ESTIMATION OF SYNTHETIC ESTIMATOR OF MEAN

In order to present some possible applications of the sequential estimation of the mean with bootstrap variance estimation at every sequential step, a population of 60 000 elements and its subpopulation of 3000 elements are generated in the following way:

1. We generate $N_1 = 20\,000$ values according to the $N(4, 1)$ distribution; we get values $x_1, \ldots, x_{2000}$ and first $k_1 = 1000$ values are transformed following the formula: $\tilde{x}_l = x_l + \varepsilon_l$, where ε_l is generated from the $N(1, 3)$ distribution. The elements $\tilde{x}_1, \ldots, \tilde{x}_{1000}, x_{1001}, \ldots, x_{20000}$ constitute the first layer of the population and the elements $\tilde{x}_1, \ldots, \tilde{x}_{1000}$ are the first small area layer.

2. We generate $N_2 = 20\,000$ values according to the $N(6, 2)$ distribution and we get values $x_{20001}, \ldots, x_{40000}$ and first $k_2 = 1000$ values are transformed following the formula from previous point. The elements $\tilde{x}_{20001}, \ldots, \tilde{x}_{21000}, x_{21001}, x_{40000}$ constitute the second layer of the population and elements $\tilde{x}_{20001}, \ldots, \tilde{x}_{21000}$ are the second small area layer.

3. We generate $N_3 = 20\,000$ values according to the $N(8, 3)$ distribution and we get values, $x_{40001}, \ldots, x_{600000}$ and first $k_3 = 1000$ are transformed following the formula from point 1. The elements $\tilde{x}_{40001}, \ldots, \tilde{x}_{41000}, x_{41001}, \ldots, x_{60000}$ constitute the third layer of the population and elements $\tilde{x}_{40001}, \ldots, \tilde{x}_{41000}$ are the third small area layer.

4. We arrange sequence $y_1, \ldots, y_{60000}$ following the formula $y_l = 3x_l + \xi_l$, where ξ_l are generated from the $N(0, \sigma)$ distribution for $\sigma = 1, 3, 5, 7$.

5. From the whole population we draw dependently a sample of size 1000 and we determine the subpopulation mean estimator value given by formula (12).

6. From the drawn sample we generate 1000 bootstrap samples and we assess the variance of estimator (12) with formula (11).

7. The variance value we got is compared with fixed value $\varepsilon^2$and we conclude the procedure or we draw 10 new elements from the population and we start all over again from point 5.

The estimates of the mean of the subpopulation considered, computed with the help of sequential estimation with bootstrap estimation of variance, are presented in Table 1.

Table 1. The sizes of samples for sequential subpopulation mean estimation for fixed precisions ε

Number of experiment	Standard deviation σ	Value of ε	Value of estimator $\hat{\theta}_0^n$	$\lvert\hat{\theta}_0^n - \theta_0\rvert$	Sample size
1	1	0.03	7.0636	0.0281	1000
2		0.01	7.0381	0.0026	1500
3	3	0.06	7.0823	0.0468	1000
4		0.03	7.0536	0.0284	1540
5		0.01	7.0423	0.0068	3750
6	5	0.09	7.1178	0.0823	1000
7		0.06	7.0399	0.0533	1190
8		0.03	7.0721	0.0366	4160
9	7	0.09	7.1342	0.0987	1000
10		0.08	7.0928	0.0773	1180
11		0.06	7.0641	0.0586	3230

Source: Author's calculations.

The actual value of θ_0 was 7.0355. In most of experiments carried out (apart from experiment 8 and 9) an estimate of parameter actual value θ_0 with accuracy not exceeding a fixed value was received. This means that the use of bootstraps to assess the variance of the mean estimator used, was successful in the cases analysed. The sequential sample size was strictly connected with the prefixed accuracy of estimation and, obviously, it grew with the growing estimation accuracy.

VI. FINAL REMARKS

The estimation of the mean, with the help of sequential methods, is connected with establishing a criterion of stopping the sequential sampling and, in consequence, with the variance or the mean square error of the estimator applied.

In the paper some simulation methods of variance estimation, among other, of the estimators of the mean, that can used in sequential estimation were presented. Particular attention was devoted the bootstrap method. This method was used to estimate the variance of the synthetic estimator of subpopulation mean. In the cases studied, the use of the bootstrap method in sequential estimation led to the estimates of subpopulation mean with precision not exceeding a prefixed number.

REFERENCES

Bracha Cz. (1998), *Metoda reprezentacyjna w badaniach opinii publicznej i marketingu*, Wyd. Efekt, Warszawa.

Dol W. (1991), *Small Area Estimation. A Synthesis between Sampling Theory and Econometrics*, Wolters Noordhoff Groningen.

Sen P.K. (1984), Nonparametric sequential estimation, [in:] *Handbook of Statistics*, vol. 4, Elsevier Science Publishers, 487–514.

Dorota Pekasiewicz

ZASTOSOWANIE METOD SYMULACYJNYCH DO SZACOWANIA WARIANCJI SEKWENCYJNEGO ESTYMATORA NIEPARAMETRYCZNEGO ŚREDNIEJ

Streszczenie

Nieparametryczne metody estymacji sekwencyjnej pozwalają, przy różnych schematach losowania próby, oszacować nieznany parametr rozkładu zmiennej losowej, gdy klasa rozkładu tej zmiennej jest nieznana.

Sekwencyjna estymacja punktowa średniej zmiennej losowej polega na wyznaczeniu wartości estymatora średniej na podstawie próby losowej, której liczebność jest odpowiednio zwiększana tak, aby funkcja ryzyka osiągnęła minimum. Jeśli nie uwzględniamy kosztów związanych z pobieraniem elementów do próby, to funkcja ryzyka jest równa błędowi średniokwadratowemu, a w przypadku estymatorów nieobciążonych wariancji stosowanego estymatora.

Wyznaczenie wariancji estymatora szacowanego parametru nie zawsze jest łatwe, a czasami nawet okazuje się niemożliwe. W statystyce małych obszarów często stosuje się estymatory pośrednie, które są bardziej efektywne niż bezpośrednie, ale ich skomplikowana postać sprawia, że często nie mamy informacji ani o ich wariancji, ani o estymatorze wariancji (lub błędzie średniokwadratowym). Przy zastosowaniu tego typu estymatorów w estymacji sekwencyjnej średniej pojawia się problem ze sformułowaniem procedury zatrzymania procesu powiększania próby. W pracy proponowane jest stosowanie, w takich przypadkach, symulacyjnych metod szacowania wariancji, m.in. metody Mahalanobisa, jackknife i metody bootstrapowej. Ponadto w pracy przedstawiony jest przykład zastosowania metody bootstrapowej do szacowania wariancji syntetycznego estymatora średniej dla podpopulacji.