Proceedings of Q2006
European Conference on Quality in Survey Statistics

Estimation of Standard Error of Indices in the Sampling Business Surveys

Rudi Seljak[1]

Abstract

Key words: standard error, indices, analytical approach, replication methods

Different types of indices are the most common output of the business short-term surveys. If these surveys are carried out on the basis of a random sample, all the estimates, calculated out of the sample data, contain sampling error. To have complete information on the quality and reliability of the published results it is necessary to estimate these sampling errors. There are two aspects of the indices as a statistical estimator that makes this task quite demanding:
Index is a non-linear estimator and in the combination with the complex sample design we have by definition difficult situation for the variance estimation.
Index is a ratio of two quantities estimated in two different time points, generally on the basis of two different samples. This fact is especially problematic when we want to use the standardised software for variance estimation since these applications are focused on the estimates based on a single sample.

A lot of different approaches and various methods for the variance estimation have been developed especially for the case of non-linear estimation and complex sample design. In the paper we present the results of the comparison study of two approaches for the estimation of standard error of indices. The first method uses analytical approach where the taylor-linearization formulas for the estimation of the ratio are adopted for the specific of the estimation of indices. The second approach uses Jack-knife replication method, again adopted for the case of estimation of indices. In the study we will use the database of monthly tax authorities' data, which will be considered as a sampling frame.

1. Introduction

The concept of index numbers is one of the widest used concepts in economic statistics, especially in the area of short-term statistics. Although the concept seems straightforward at the first sight, the theory of index numbers faces a lot of difficulties in the process of the implementation of the theoretical concepts into statistical and economic reality. The reason for the pretentiousness of the index numbers theory and practice comes from the fact that we usually use the index numbers to describe complex and dynamic economic phenomena and doing it we use the statistical tools like sampling and estimation procedures which are pretentious themselves. Hence

[1] Rudi Seljak, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia, rudi.seljak@gov.si, phone: +386 1 2415 126

we can consider two general aspects of the concepts of index numbers, the formal-mathematical one and the practical-economic one. The real challenge in index theory is to build the bridge between both aspects and consequently provide the tool which would be theoretically correct and would at the same time serve for the purposes of clear and useful description of the economic situation.

The paper focuses on indices which are calculated out of the sampling survey data. By the definition all the estimates from such a survey contain sampling error and the estimation of these errors is the common task for survey statisticians. Estimations of indices are usually done on the basis of the panel survey data where the same sample is surveyed in consequent time intervals. Due to the practical reasons we don't constantly include the same units in the sample, but we rotate part of the units out of the sample and replace them with new ones in regular time intervals. The consequence of such a procedure is that some of the indices are estimated from two different (partly overlapping) samples where two different grossing-up weights are used. This fact along with the fact that index is a ratio, hence a non-linear estimator, makes the task of variance estimation of indices quite demanding.

The main goal of the paper is to describe the first results of the simulation study which was carried out at the Statistical Office of the Republic of Slovenia (SORS). The study was focused on some of the properties of the index as a statistical estimate obtained from the sampling survey data. The main goal of the study was to explore the sampling distribution of the index estimator and to compare three different methods for estimation of the sampling variance and the standard error of the estimate.

In the first part of the paper we will briefly introduce some basic concepts from the index and from the sampling theory. Then we will describe the simulation study with emphasis on the description of all three methods for variance estimation. We will then show the main results of the study and at the end some conclusions will be given. We have to stress that the study is not finished yet and the presented results are just the results of the first stage of the study. In the future we plan to broaden the field of the study to some other types of indices and some of the results of the first part of the study should be explored more detailed.

2. Basic theory

2.1. Index numbers

Index is generally not exclusively economic concept but is also used in some other areas as for instance in the social studies. However in our paper index will only be considered as a concept referring to the economic theory and practice. Generally index is defined as a ratio of two or more values all measured with the same unit. Index can compare values in time or space but for the purposes of this study we will only consider time indices, hence we define index as a ratio of two or more values, measured with the same unit in two different time points. There are many different approaches and out of these approaches derived formulas for constructing an index.

The basic approach however is built on the assumption that we are observing pairs of price-quantity for basket of n different commodities in two different time points. The index which is derived from this approach is called the aggregative index. The general form of aggregative price index is

$$P_{t0} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_i}{\sum_{i=1}^{n} p_{i0} \cdot q_i} \quad (2.1)$$

Here the quantities q_i play the role of weights, causing that the commodities with higher purchase have more significant influence to the value of the index. In the above general formula the time period of the quantities q_i is still undefined. The two most widely used approaches are to take the weights from current time point t or from the base time point 0. The former case leads us to the Paasche price index and the later one to the Laspeyres price index.

Sometimes it is however difficult or even impossible to observe separately quantities and prices and in this case we have to deal with the value index:

$$V_{t0} = \frac{\sum_{i=1}^{n} p_{it} \cdot q_{it}}{\sum_{i=1}^{n} p_{i0} \cdot q_{i0}} = \frac{V_t}{V_0} \quad (2.2)$$

In this case there are no weights used therefore we don't have to distinguish different forms of the index according to the time period of the weights. One of the special cases of the value index is the turnover index which will be the primer concern of our paper. Let us imagine an economic system where in the current time point t we have N_t units which created turnover $T_t = \sum_{i=1}^{N_t} T_{it}$ whereas in the base time period 0 the system consisted of N_0 units which created the turnover $T_0 = \sum_{i=1}^{N_0} T_{i0}$. Turnover index is then defined as

$$T_{t0} = \frac{T_t}{T_0} \quad (2.3)$$

Very often it is the case that we are not able to observe all of the N_t (N_0) units and we are observing just the selected random sample of n_t (n_0) units. In this case the total turnover for each of the periods is estimated by using the grossing up weights w_{it} (w_{i0}). The turnover index of current period t compared to the base period 0 is then estimated by:

$$\hat{T}_{t0} = \frac{\sum_{i=1}^{n_{tt}} w_{it} \cdot T_{it}}{\sum_{i=1}^{n_0} w_{i0} \cdot T_{i0}} = \frac{\hat{T}_t}{\hat{T}_0} \quad (2.4)$$

The standard error of this estimator will be the main subject of our paper.

2.2. Sampling error

Most of the statistical results, especially in the area of short-term statistics, are the estimates, calculated on the basis of a random sample. The consequence of the fact that that we are observing just a part of the target population is that all these results contain sampling error. To enable correct interpretation of such results it is necessary to produce and publish also the estimates of the sampling error of the published results. Many different methods and techniques have been developed in order to achieve this goal. All these methods could roughly be divided by two basic approaches: analytical approach and re-sampling approach.

Using the analytical approach we try to find exact or at least approximate formula for (at least approximately) unbiased estimator of sampling variance. The exact formulas are more or less reserved for the case of linear estimators whereas in the case of non-linear estimators we usually have to deal with approximate formulas. Most of these approximate formulas are derived by using the mathematical technique called Taylor series linearization. The index estimator is the special case of so called estimation of the ratio, which has the mathematical form: $\hat{R} = \frac{\hat{X}}{\hat{Y}}$, where $\hat{X}$ and $\hat{Y}$ are estimates of the population totals. The approximate formula for the sampling variance of the estimation of the ratio is:

$$\mathrm{var}(\hat{R}) = \frac{1}{\hat{Y}^2}\left[\mathrm{var}(\hat{Y}) + \hat{R}^2 \cdot \mathrm{var}(\hat{X}) - 2\hat{R} \cdot \mathrm{cov}(\hat{Y}, \hat{X})\right]$$

(2.5)

This formula will be the basis for the analytical approach of the variance estimation of the index estimator.

The re-sampling approach for variance estimation is more practical than theoretical approach. The basic idea is that we use initial sample to produce many subsamples, estimate the parameter of interest from each of these subsamples and than use the variability of the produced estimates to estimate the sampling error of the initial estimate. There are many variations of this idea, all of them of course based on the strict mathematical theory. We will use the method which is known as the Jackknife (JKK) method, more precisely the JKK-1 method. The idea here is to form the subsamples by deleting each time one of the units from the initial sample. If $\hat{\theta}$ is the estimate of the parameter of interest produced from the initial sample, $\hat{\theta}_i$ estimates from the subsamples and k number of subsamples than the formula for variance estimation is as follows:

$$\mathrm{var}(\hat{\theta}) = \frac{1}{k(k-1)} \sum_{i=1}^{k} (\hat{\theta} - \hat{\theta}_i)^2 \quad .$$

(2.6)

3. Simulation study

3.1. Target population and sample design

The basis for the simulation study was the monthly data from the TAX authorities for the enterprises from sections G-I of NACE classification. Out of the TAX data we were able to get he good approximation of the monthly turnover for approximately 18 000 enterprises for each month of years 2004 and 2005. These enterprises were then considered as our target population which was slightly changing from month to month. Out of this population we simulated the sampling survey procedure in order to obtain estimate of several different turnover indices, calculating regarding to different time periods.

Several different variations of sampling design were used during the simulation study but the basic features of the sampling procedure were always the same. When we choose these basic features we tried to follow the usual practice in the surveys whose main goal is the estimation of the indices. The initial sample which was selected for the reference period May 2004 was stratified simple random sample. The stratification was done according to activity group and size class. To determine the activity group we used the 2-digit level of NACE classification. The three size groups were determined on the basis of the annual turnover of the enterprises which was determined by the historical TAX data for each of the enterprises in the target population. Since it is known that the panel survey is the most appropriate for the purpose of estimation of indices, we used such a design in our study. Therefore for several months after May 2004 we kept the same (survived) units in the sample and after the particular time period we refreshed the sampling frame, rotated part of the units out of the sample, replaced them with the new units and then again used the same sample for the particular time period. The coordination of the samples was assured by using the well known permanent random numbers system. The sample size, the rotation rate and the time gap between the two consequence rotations were the subject of the parameterization of the simulation. The aim of the first part of the study was to explore the sampling distribution of the estimator of the chosen index. In order to get the approximate sampling distribution we repeated the whole sampling procedure for the chosen parameterization for 10 000 times, estimated the index from each of the chosen samples and so we have got 10 000 estimates of the index. The histogram of these estimates together with some basic statistics was the basic output of the first part of the study.

The second part of the study was devoted to the research and comparison of three different methods (which will be explained later) for the estimation of the sampling variance which by definition equals to the variance of the sampling distribution. The procedure of variance estimation is much more time consuming than just the estimation of the parameter of interest that why we couldn’t afford to repeat the procedure for 10 000 times, but we have to limit the procedure to 1000 or 2000 repetitions. In this case after each of the sample selection we estimated the sampling variance of the estimated index by each of the three methods and so we have got 1000 estimates for the sampling variance, produced by each of the methods. The main goal was to compare biasness and variance of each of the methods.

3.2. Methods for variance estimation

The first two tested methods are in fact just the variations of the same concept, based on the formula (2.5). The estimates $\hat{T}_t$ and $\hat{T}_0$ in formula (2.4) are generally estimated from two (at least partly) different samples where for the same unit in denominator and enumerator two different weights could be used. This fact causes departure from the classical problem of the variance estimation of the estimated ratio of two totals. Since variance of $\hat{T}_t$ and $\hat{T}_0$ could easily be estimated separately, the real problem is just how to estimate the covariance $\mathrm{cov}(\hat{T}_t, \hat{T}_0)$. The first two methods in fact differ just in the way how this covariance is estimated.

For the first method we tried to use the existed possibilities of SAS software in order to avoid the redundant programming as much as possible. We used SAS procedure SURVEYMEANS for the estimation of $\mathrm{var}(\hat{T}_t)$ and $\mathrm{var}(\hat{T}_0)$, but the procedure doesn't provide the estimator of sampling covariance. That why we used the well known relation:

$$\mathrm{cov}(\hat{T}_t, \hat{T}_0) = \frac{1}{2}\left[\mathrm{var}(\hat{T}_t) + \mathrm{var}(\hat{T}_0) - \mathrm{var}(\hat{T}_t + \hat{T}_0)\right]$$

and transform formula (2.5) as follows:

$$\mathrm{var}(\hat{T}_{t0}) = \mathrm{var}\left(\frac{\hat{T}_t}{\hat{T}_0}\right) = \frac{1}{\hat{T}_0^{\,2}}\left[(1+\hat{T}_{t0})\cdot\mathrm{var}(\hat{T}_t) + (\hat{T}_{t0} + \hat{T}_{t0}^{\,2})\cdot\mathrm{var}(\hat{T}_0) - \hat{T}_{t0}\cdot\mathrm{var}(\hat{T}_t + \hat{T}_0)\right]$$

(3.1)

To correctly estimate the $\mathrm{var}(\hat{T}_t + \hat{T}_0)$, we had to "construct" the common weight for the sum. We used the equality $w_{i,sum}(T_{it} + T_{i0}) = w_{it}T_{it} + w_{i0}T_{i0}$ to get:

$$w_{i,sum} = \frac{w_{it}T_{it} + w_{i0}T_{i0}}{T_{it} + T_{i0}}$$

(3.2)

Later in the paper we will refer to this method by using the abbreviation AN1.

The second method is also based on the analytical approach, using the same formula (2.5) as the first method. The only difference with regarding to the first method is that we here used the direct formula for the estimation of the sampling covariance. The formula for the case of stratified simple random sampling without replacement which takes into account the chancing of the population, has been derived in [9]:

$$\mathrm{cov}(\hat{T}_t, \hat{T}_0) = \sum_{h=1}^{H}\left(1 - \frac{n_h(t)n_h(0)}{N_h(t)N_h(0)}\frac{N_h(t,0)}{n_h(t,0)}\right)\cdot\frac{n_h(t,0)}{n_h(t,0)-1}\sum_{k=1}^{n_h(t,0)}(z_{hk}(t) - \bar{z}(t))(z_{hk}(0) - \bar{z}(0))$$

(3.3)

$N_h(t)$ .....number of population units in stratum h in time point t

$n_h(t)$ .....number of sample units in stratum h in time point t

$N_h(t,0)$ ...number of population units in stratum h in both time points t and 0

$n_h(t,0)$ ....number of sample units in stratum h in both time points t and 0

$z_{hk}(t) = \frac{N_h(t)}{n_h(t)} T_{hk}(t)$

$\bar{z}(t)$ …….average of the elements $z_{hk}(t)$

We will denote this method with the abbreviation AN2.

The third method that we tested was the well known Jackknife re-sampling method where the replications are created by deleting each time one unit from each stratum. In the concrete realization of the method we used SUDAAN software to perform the method. To enable the usage of the already programmed procedure inside the SUDAAN software we first had to adjust our data. Since the procedure presumes just one sample with unique weight for each of the units, we merged the two samples for time points t and 0 into one data set, putting missing values to 0. For the units that have been included in both samples we constructed the artificial unique weight by simple taking the average of both of the weights from different samples. Method performed in such a way is not completely correct realization of the JKK method for the case of estimated indices, but we estimated that the above described simplifications shouldn't influence the results significantly. The tailor made programming of totally correct execution of the method would simply demand too much work and we estimated that the gains that we would get wouldn't compensate for this input. For this method we will use the abbreviation JKK.

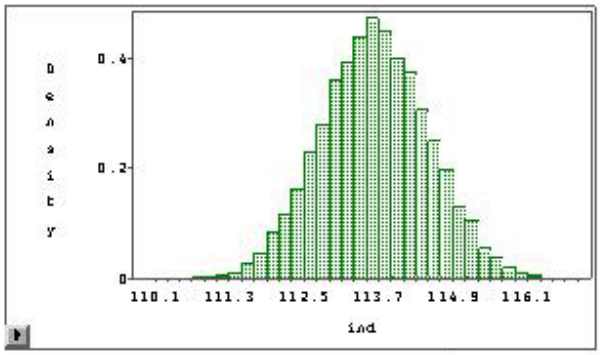
4. Results of the study

4.1. Sampling distribution

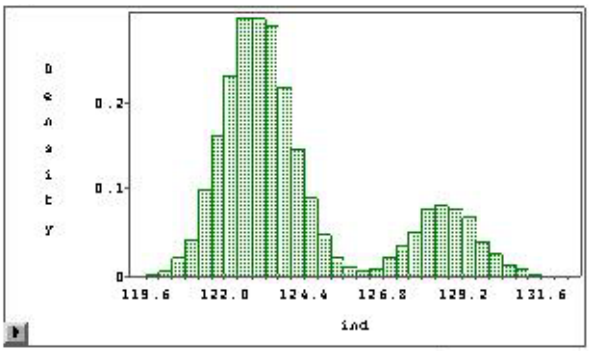
The first aim of the study was to explore the sampling distribution of the index-estimator for different variations of the basic sampling design explained above. According to the general theory, the distribution of the index estimator, which is just the special case of the estimator of the ratio, should approximately be normal. The main goal of this part of the study was hence most of all to check this assumption in the case of real data and to find eventual cases of significant departures from the normal distribution.

In order to be able to get the sufficient precision of the picture of the sampling distribution, the stochastic procedure for each of the variations of the sampling design was repeated for 10000 times. For each of the selected pairs of the time points we hence had 100000 weighted sums of turnover and therefore 100000 estimates of turnover index. By plotting the histogram of these estimates we have got quite a clear picture of the distribution of the index-estimator.

In most of the cases the distribution of the index was indeed quite clear (approximately) normal and the plotted histogram looked like the one presented in the picture below:

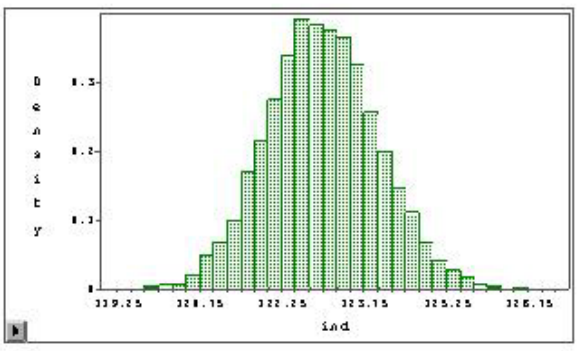


But in some cases the shape of normality disappeared and we have been faced with clear bimodal distribution:



4.2. The problem of bimodality

The above presented bimodal distribution is the consequence of the used sampling design when all the large units were sampled with certainty. The reason for the distort distribution is in fact just one unit which in the base time period was stratified as a medium enterprise but has later significantly increased its turnover. If we would look all the values in the population this value wouldn't be considered as an outlier but if we remove all the large units which are included in the sample with certainty, this unit becomes an obvious outlier. We will call such unit the hidden outlier. If we remove the detected hidden outlier from the population and repeat the simulation procedure, we get the following sampling distribution:

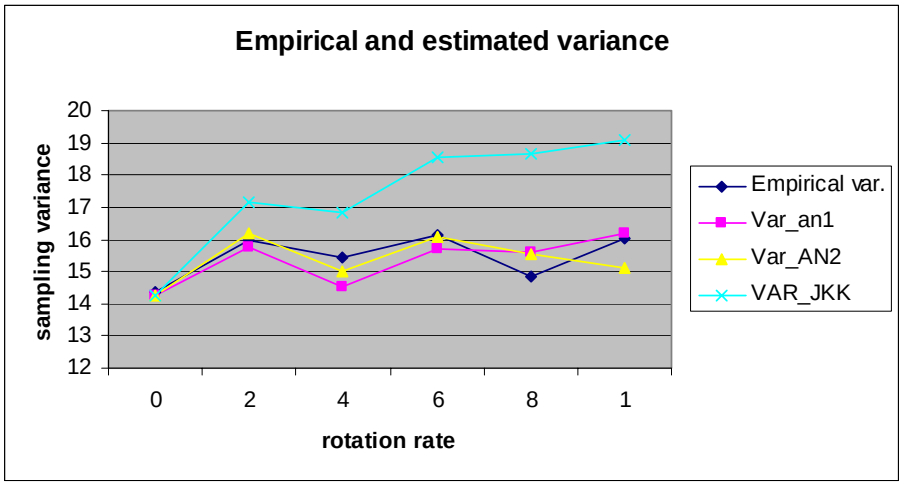


Comparing the standard deviation of both of the distribution we see that in the case of the bimodal distribution the standard deviation is 2.5 while in the case where the

hidden outlier was removed and the distribution was approximately normal, the standard deviation is just 1. Hence just one outlying value caused the increase of the sampling error for the factor 2.5. Beside the significant enlargement of the sampling variance, bimodal distribution also means that in this case the confidence interval could not be interpreted as in the case of approximately normal distribution. In fact in this case the confidence interval doesn't provide much of the information at all. But how can we avoid the situation of bimodal distribution. Probably the only way is to detect the hidden outlying values in the editing process and then treating them as a large units meaning that they become self-representative units with sampling weight 1. All that was said emphasis again the importance of efficient editing procedure for the quality of the final results.

4.3. Variance estimation

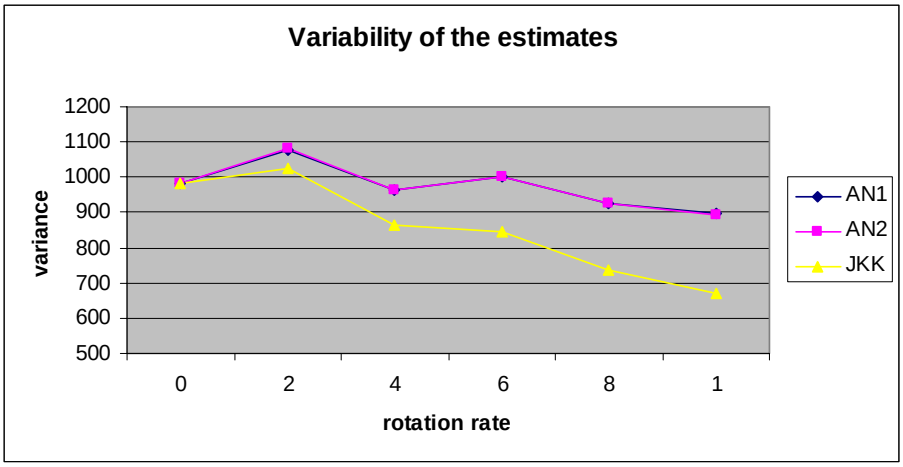
We will separately treat the case of normal and the case of bimodal distribution. In the case when the sampling distribution is normal, all three methods perform quite well. In the graphs that follow we show the comparison of empirical sampling variance (derived out of the 10000 repetitions of the procedure) with the average[2] of the 10000 estimates by each of the above described methods for variance estimation. We show the case when we fixed both time points and the sample size and we were changing just the rotation rate of the sample for the second time point.



As we can see from the graph, the estimates obtained on the basis of the analytical methods are quite close to the "true" sampling variance, whilst the Jackknife estimates overestimates the variance, especially in the cases of higher rotation rates. Here it should be mentioned that the overestimation of the JKK method is not due to the method itself, but rather due to some simplifications in the process of the adjustment of the method for the case of index estimation. Therefore this bias could be decreased by some additional work on the "technical" details of the application of the method.

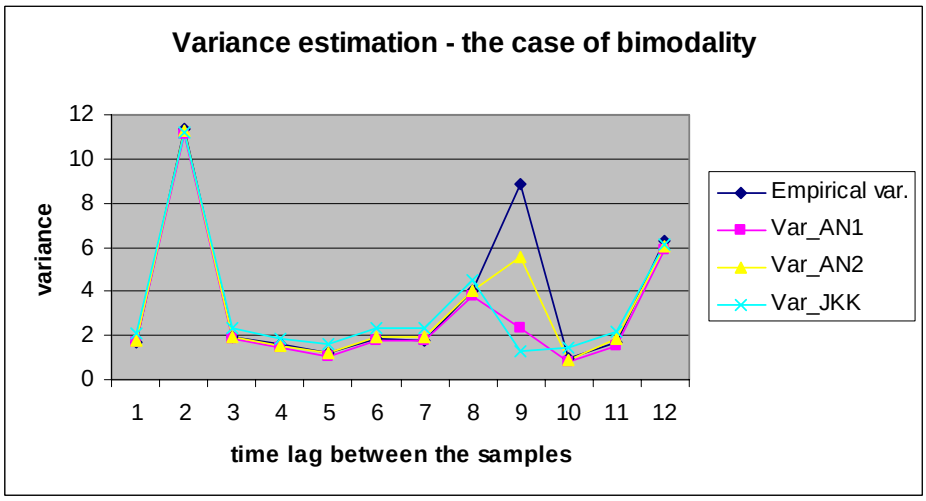
[2] With 10000 repetitions we can consider this average quite good estimation for the theoretical concept of the expected value of the variance estimator

Another important aspect of each of the statistical estimates is its variability, usually expressed in the form of the variance. In the next picture we present the variance of the variance estimation, obtained by all of the three methods. The variance is the measure of dispersion of the estimates and could also be consider as the indicator of the stability of the method.



The visual impression that there are just two methods presented in the graph is due to the fact that the variances of the analytical methods are almost the same. It is quite clear that, especially in the cases of higher rotation rates, by this indicator the JKK method performs better than the analytical methods.

At the end let us present how the methods perform in the case of the bimodal distribution. We will present the results for the case when the sample size (4000) and the rotation rate (0.2) were fixed and we were changing the time lag (in months) between the two time points of comparison. As it will be clearly seen from the pictures, the case of bimodality occurs twice. We first present the comparison of the averages of the estimates compared to the empirically determined sampling variance.



We can see that in the first case of bimodality, all the methods “follow” the leap of the sampling variance, whereas in the second case, the methods perform quite

differently. So far we couldn't find the theoretical justification for such a behavior and it should be a subject of further investigations.

5. Conclusions

In the paper we presented the first results the simulation study which is carried out on the basis of the large dataset of tax authorities' data. Out of these data we were able to get quite good approximation for the monthly turnover for large amount of the enterprises from sections G-I of NACE classification. The set of enterprises with the belonging data on turnover was considered as a sampling frame in the simulation study.

The first part of the study was devoted to the research of the sampling distribution of the index estimator. The sampling design was, as usual in the case of the index estimation, rotating panel sample where in the first month stratified simple random sample was selected and in the second month the part of the selected sample was replaced with the new units. In both months all the large enterprises were selected with certainty. To get a clear picture of the sampling distribution, the stochastic procedure of the sample selection and index estimation was repeated for 10.000 times. The sampling distribution was in most of the cases approximately normal what is to be expected according to the sampling theory. But in some cases we have been faced with quite clear bimodal distribution. The reason for the bimodal distribution are so called "hidden outliers", these are the enterprises which were initially selected in the sample as medium enterprises but have significantly increased their turnover in the second month. These enterprises should be treated as outliers just in the case of chosen sampling design where all the large enterprises were selected with certainty. If for instance the sampling design would be SRS, these units wouldn't be outliers and the sampling distribution wouldn't be bimodal. The more exact mathematical model which would explain and forecast the bimodal distribution should be developed in the future.

The second aim of the study was to find the most efficient method for the estimation of the sampling variance of index estimator[3]. For this purpose three different methods were tested, two of them based on the analytical and one on the re-sampling approach. The main findings of the study could be summarized as follows:

- The second analytical method performs slightly better but the advantage of the first method is that it requires less tailor made programming.
- The Jackknife method slightly overestimates the variance but we judge this bias is due to the technical reasons of adjustment of the method and it could be decreased.
- The variability of the estimates is the lowest in the case of JKK method.
- The problem of bimodality is the problem which should be further investigated in the future.
- Bimodal sampling distribution can cause serious inconsistency in the procedure of variance estimation.

[3] In fact the variance estimation is the primer intention of the study. Research of the sampling distribution, especially bimodality, became subject of the study later through the first results.

- Bimodality of the distribution should require different interpretation of the sampling variation.

In the future the study will focus on some other types of indices. Special concern should be devoted to the chained indices which are very often used in the economy statistics. Also the concrete applications for the variance estimation of different types of indices should be developed.

References

[1] Allen, R.G.D. (1975) Index Numbers in Theory and Practice. Chicago: Aldine.
[2] Brenda G. Cox et al. – Business Survey Methods, Wiley, 1995
[3] Kish Lesley: Survey sampling:John Wiley & sons, 1965
[4] L. Lyberg et al. – Survey measurement and Survey Quality, Wiley, 1997
[5] Mick Silver:Business Statistics:The McGraw-Hull Companies, 1997
[6] Methodological documents, Standard Report: Working Group "Assessment of quality in statistics", Sixth meeting, Luxembourg 2-3 October 2003
[7] Methodological documents, Definition of Quality in Statistics: Working Group "Assessment of quality in statistics", Sixth meeting, Luxembourg 2-3 October 2003
[8] Sarndal Carl-Erik: Model Assisted Survey Sampling: Springer-Verlag, 1992
[9] Tam, S.M. (1984) - On Covariances from Overlapping Samples; The American Statistician, 38, pp. 288-292