ANADOLU ÜNİVERSİTESİ

Bilim ve Teknoloji Dergisi A-Uygulamalı Bilimler ve Mühendislik
Cilt: 14 Sayı: 3 2013
Sayfa: 231-239

ARAŞTIRMA MAKALESİ /**RESEARCH ARTICLE**

Barbara KOWALCZYK[1]

ON THE USE OF AUXILIARY INFORMATION IN MULTIPURPOSE ROTATING SURVEYS

ABSTRACT

In most repeated surveys interest centers both on the estimation of the population mean on each successive occasion and estimation of further population parameters like net changes, number index estimation or average over several occasions. Due to this fact many of repeated surveys are rotating surveys. The multipurpose nature of repeated surveys poses additional questions to the already difficult theory of rotating surveys. The aim of the paper is to analyze how the use of auxiliary information on each successive occasion and consequently increase of the efficiency of the mean estimation on each occasion affects efficiency of net changes estimation in the case of repeated rotating surveys.

Keywords*:* Repeated rotating surveys, Multipurpose surveys, Auxiliary information

YARDIMCI BİLGİNİN ÇOK AMAÇLI DÖNEN ANKETLERDE KULLANIMI

ÖZ

Tekrarlı birçok ankette ilgi ardışık her bir olayın anakütle ortalamasının tahmini ve net değişim, sayı indeks tahmini veya belirli olay ortalamaları gibi diğer anakütle parametrelerinin tahmini üzerine yoğunlaşmaktadır. Bu nedenle tekrarlı anketlerin birçoğu dönen anketlerdir. Tekrarlı anketlerin çok–amaçlı yapısı zaten oldukça karmaşık olan dönen anketler teorisine birçok soruyu daha eklemektedir. Bu çalışmanın amacı ardışık her olay için yardımcı bilgi kullanımının ve sonuç olarak her bir olayın ortalama tahmininin etkinliğindeki artışın net değişim tahminini nasıl etkileyeceğini tekrarlı dönen anketler ile analiz etmektir.

Anahtar Kelimeler*:* Tekrarlı dönen anketler, Çok amaçlı anketler, Yardımcı bilgi

[1] Warsaw School of Economics, Institute od Econometrics, Al. Niepodleglosci 162, Warszawa 02-554; Poland
Tel: 48 22 564 9258, E-mail: bkowal@sgh.waw.pl

Recieved: 28 November 2012 **Revised:** 16 April 2013 **Accepted:** 21 June 2013

1. INTRODUCTION

Various statistical problems connected with repeated rotating surveys have constituted important issue among statisticians from years, see e.g. Binder and Dick (1989), Patterson (1950), Duncan and Kalton (1987), Jones (1980), Rao and Graham (1964) and still play important role in modern statistics, see e.g. Berger and Priam (2010), Kim et al. (2005), Kordos (2012), Kowalczyk (2003), (2004), (2013), Kowalski (2009), McLaren and Steel (2001), Park et al. (2001), Wesołowski (2010). In the present paper problem connected with multipurpose nature of repeated rotating surveys is considered.

In finite population studies majority of repeated surveys are of multipurpose nature, which means that on the basis of one repeated survey we usually want to estimate several population parameters, like population mean or total on successive occasions, net changes between two successive occasions, population ratio, average of means or total over several occasions etc. Due to this fact many of repeated surveys are rotating surveys, which means that sample on each occasion consists of two parts: elements that has been examined also on previous occasion and elements that are new in the sample, that is elements that hasn't been examined on previous occasion. The problem that is considered in the paper regarding rotating surveys is the following. We focus our attention on the question how increase of the efficiency of mean estimation affects estimation of net changes, that is changes between population means on two successive occasions. When the main aim of the survey is to estimate population mean on each successive occasion we want to increase efficiency of the mean estimation, which we often do in finite population surveys by using auxiliary information and composite estimators, e.g. difference estimators. When on occasion t, to estimate population mean, we use difference estimator, which takes into account auxiliary information and on occasion t+1, to estimate population mean, we also use difference estimator, which take into account auxiliary information we consequently have to use the same estimators to estimate net changes. Influences of such a procedure on properties of net changes estimator are presented in the paper.

Formally situation described below is as follows. A simple random sample without replacements SRSWR of n elements is drawn from a finite population of N-elements on the occasion t. A sample on the occasion t+1consists of n elements, from which np, $0 \le p \le 1$, elements are retained from the previous sample, that is selected by SRSWR from n elements that have been examined on occasion t and the remaining n-np elements are drawn by SRSWR from the population elements N-n that has not been examined on the previous occasion. We assume here that on two successive occasions structure of the population does not change, that is $N_t = N_{t+1} = N$, see also e.g. Kowalczyk (2003), Patterson (1950), Wesolowski (2010).

Sample on occasion t consists of n elements from which np are examined on occasion t and will be examined on occasion t+1 (they are matched elements) and n-np elements that are examined on occasion t and will not be examined on occasion t+1 (they are unmatched elements).

Analogously, sample on occasion t+1 consists of n elements from which np elements are examined on occasion t+1 and have also been examined on occasion t (they are matched elements) and n-np elements that are examined on occasion t+1 and have not been examined on occasion t (they are unmatched elements).

On each occasion to estimate population mean difference estimators based on auxiliary information are used, more precisely on occasion t to estimate population mean we use difference estimator of the form:

$$\bar{y}_t^D = \bar{y}_t + B_t(\bar{X}_t - \bar{x}_t) \tag{1}$$

and on occasion *t*+1 to estimate population mean we use difference estimator of the form:

$$\bar{y}_{t+1}^D = \bar{y}_{t+1} + B_{t+1}(\bar{X}_{t+1} - \bar{x}_{t+1}) \tag{2}$$

Due to this fact for assessing net changes, that is for assessing $\overline{Y}_{t+1} - \overline{Y}_t$, in multipurpose surveys we consequently have to use estimator of the form:

$$d^D = \overline{y}^D_{t+1} - \overline{y}^D_t . \tag{3}$$

It is well known that difference estimators are unbiased estimators of a population mean and they increase efficiency of mean estimation compare to the usual sample mean, that is:

$$D^2(\overline{y}^D_t) = \left(1 - \frac{n}{N}\right)\frac{S^2(Y_t)}{n}(1 - \rho^2(Y_t, X_t)) \tag{4}$$

and analogously

$$D^2(\overline{y}^D_{t+1}) = \left(1 - \frac{n}{N}\right)\frac{S^2(Y_{t+1})}{n}\left(1 - \rho^2(Y_{t+1}, X_{t+1})\right) \tag{5}$$

See e.g. Särndal et all (1992).

The problem arises when additionally net changes estimation is taken into account. It has to be remembered that samples on both occasions overlap in rotating surveys, which introduces additional complexity to the problem. In the next section properties of the estimator d^D will be studied. Estimator d^D will also be compared to the usual one, i.e. to the estimator d of the form:

$$d = \overline{y}_{t+1} - \overline{y}_t \tag{6}$$

2. PROPERTIES OF THE ESTIMATOR

When analyzing net changes estimation, first important step is to compare precision of the estimator d^D given by (3), which takes into account auxiliary information with the precision of the usual estimator d given by (6).

Necessary and sufficient condition for d^D to be superior to the usual estimator d , that is necessary and sufficient condition for $D^2(d^D) \le D^2(d)$, should in some way depend on the matched fraction p in the sample. As we see later in this section, as far as auxiliary information is taken into account, the following formula proves to play important role in the analysis:

$$\begin{aligned} M = {} & \rho(Y_{t+1}, Y_t) - \rho(Y_{t+1}, X_t)\rho(Y_t, X_t) - \rho(Y_{t+1}, X_{t+1})\rho(Y_t, X_{t+1}) + \\ & + \rho(Y_{t+1}, X_{t+1})\rho(Y_t, X_t)\rho(X_{t+1}, X_t) \end{aligned} \tag{7}$$

Exact conditions for $D^2(d^D) \le D^2(d)$ to be satisfied are given in theorem 1.

Theorem 1: If $\rho(Y_{t+1}, Y_t) - M > 0$, where M given by (7), necessary and sufficient condition for d^D to be superior to the usual estimator d is the condition for the matched fraction p in the sample:

$$p \le \left(\frac{1}{2}\left(1 - \frac{n}{N}\right)\frac{\frac{S(Y_{t+1})}{S(Y_t)}\rho^2(Y_{t+1}, X_{t+1}) + \frac{S(Y_t)}{S(Y_{t+1})}\rho^2(Y_t, X_t)}{\rho(Y_{t+1}, Y_t) - M} + \frac{n}{N} \right). \tag{8}$$

If $\rho(Y_{t+1}, Y_t) - M < 0$, where *M* given by (7), necessary and sufficient condition for d^D to be superior to the usual estimator d is the condition for the matched fraction *p* in the sample:

$$p \geq \left(\frac{1}{2}\left(1 - \frac{n}{N}\right) \frac{\frac{S(Y_{t+1})}{S(Y_t)}\rho^2(Y_{t+1}, X_{t+1}) + \frac{S(Y_t)}{S(Y_{t+1})}\rho^2(Y_t, X_t)}{\rho(Y_{t+1}, Y_t) - M} + \frac{n}{N} \right). \tag{9}$$

Proof: see Kowalczyk (2013).

The next important question that arises in multipurpose rotating surveys in which auxiliary information is used is the following. It is well known that to maximize efficiency of the estimation for estimating net changes, when using usual estimator $d = \bar{y}_{t+1} - \bar{y}_t$, it is best to retain the same sample throughout all occasions providing that the correlation of Y_{t+1} and Y_t is positive. Otherwise, that is if the correlation of Y_{t+1} and Y_t is negative, it is best to select a new sample for the next occasion. When not only two study variables but also auxiliary variables are taken into account the problem becomes much more complicated. So the question arises when we should retain the same sample for the next occasion and when we should select a new sample if we want to maximize efficiency of the net changes estimation. Theorem 2 gives answer to this question.

Theorem 2: If we assess net changes by using composite estimator of the form d^D given by (3) we have the following:

if $M > 0$, where *M* is given by (7), minimum $D^2(d^D)$ is achieved for matched fraction $p = 1$;
if $M < 0$ where *M* is given by (7), minimum $D^2(d^D)$ is achieved for matched fraction $p = 0$.

Proof: see Kowalczyk (2013).

In other words, theorem 2 tells us, that it is best to retain the same sample for the next occasion providing that *M* is positive, and it is best to select a new sample for the second occasion providing that *M* is negative.

It is worth to notice that when composite estimator is applied it is possible that correlation coefficient between study variable on two occasions is positive and nevertheless it is best to select a new sample for the second occasion to maximize efficiency of the net changes estimation. The answer to the question if we should retain the same sample for the next occasion or select a new sample does not depend in this case simply on the correlation coefficient between study variable on two occasions, it depends on *M* given by (7), i.e. it depends on the combination of all correlation coefficients (between study and auxiliary variables).

3. SIMULATION STUDY

3.1 Population 1

According to multivariate normal distribution a population of N=15000 elements was generated. Finite population parameters were as follows:

$$\begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} 5.0153 \\ 19.9248 \\ 39.9866 \\ 50.0008 \end{bmatrix}, \quad S = \begin{bmatrix} 2.0135 \\ 7.9497 \\ 7.9494 \\ 10.0633 \end{bmatrix}, \quad \mathrm{R} = \begin{bmatrix} 1 & 0.3980 & 0.7019 & 0.7019 \\ 0.3980 & 1 & 0.6972 & 0.7003 \\ 0.7019 & 0.6972 & 1 & 0.5027 \\ 0.7019 & 0.7003 & 0.5027 & 1 \end{bmatrix}$$

On two occasions samples of 800 elements were selected according to SRSWR; four different matched fractions of two samples were taken into account: p=0, p=0,4, p=0,8, p=1. For each case sampling was repeated h=1000 times and for each selected sample estimates based on estimator d^D given by (3) and estimates based on estimator d given by (6) were obtained. For two estimators and 1000 repetitions MSE were obtained:

$$MSE(d) = \frac{1}{1000}\sum_{j=1}^{1000}(d - (\overline{Y}_{t+1} - \overline{Y}_t))^2,$$

$$MSE(d^D) = \frac{1}{1000}\sum_{j=1}^{1000}(d^D - (\overline{Y}_{t+1} - \overline{Y}_t))^2.$$

Also percentage efficiency of the estimator d^D compared to the estimator d was obtained:

$$EFFP(d^D) = \frac{MSE(d) - MSE(d^D)}{MSE(d^D)} \cdot 100\%.$$

Results of a simulation study for population 1 are presented in Table 1 and table 2.

Table 1. Mean Square Error and percentage efficiency of the estimation

	p=0	p=0,40	p=0,80	p=1
$MSE(d)$	0,0799	0,0773	0,0716	0,0662
$MSE(d^D)$	0,0394	0,0469	0,0546	0,0535
$EFFP(d^D)$	102,8%	65,0%	31,1%	23,6%

As we can see from table 1 estimator d^D, which takes into account auxiliary information, is in all cases superior to the usual estimator d, i.e. $MSE(d^D)<MSE(d)$ for all p. For estimator d efficiency of the estimation is the highest, that is MSE is the lowest for p=1, that is for the same sample on two occasions. For estimator d^D efficiency of the estimation is the highest, that is MSE is the lowest for p=0, that is for disjoint samples on two occasions.

Now we can compare these results with theoretical results from section 2.

On the basis of theorem1 for this population we have:

- $D^2(d^D) \le D^2(d)$ for all $p \in \langle 0,1 \rangle$

On the basis of theorem2 we have:

- min $D^2(d^D)$ is achieved for p=0, because $M < 0$.

From statistical theory we have:

- min $D^2(d)$ is achieved for p=1, because $\rho(Y_2, Y_1) > 0$.

As we can see all simulation results agree with theoretical ones.

From table 2 we can see that gain in the efficiency of the estimator d^D compare to the usual estimator d can be very high, it can exceed 100%. Thus it has strong practical importance.

3.2 Population 2

According to multivariate normal distribution a population of *N*=5000 elements was generated. Finite population parameters were as follows:

$$\begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} 4{,}007 \\ 5{,}007 \\ 3{,}005 \\ 4{,}011 \end{bmatrix}, \; S = \begin{bmatrix} 0{,}809 \\ 0{,}982 \\ 0{,}602 \\ 0{,}799 \end{bmatrix}, \; R = \begin{bmatrix} 1 & 0{,}410 & 0{,}502 & 0{,}612 \\ 0{,}410 & 1 & 0{,}595 & 0{,}495 \\ 0{,}502 & 0{,}595 & 1 & 0{,}394 \\ 0{,}612 & 0{,}495 & 0{,}394 & 1 \end{bmatrix}$$

On two occasions samples of 100 elements were selected according to SRSWR; five different matched fractions of two samples were taken into account: *p*=0, *p*=0,2, *p*=0,5, *p*=0,8, *p*=1. For each case sampling was repeated *h*=1000 times.

Results of a simulation study for population 2 are presented in Table 3 and table 4.

Table 2. Mean Square Error and percentage efficiency of the estimation

	p=0	*p*=0,20	*p*=0,50	*p*=0,8	*p*=1
MSE(*d*)	0,0145	0,0143	0,0129	0,0119	0,0103
MSE(d^D)	0,0114	0,0114	0,0119	0,0139	0,0146
EFFP(d^D)	27,1%	25,4%	7,6%	-14,3%	-29,1%

As we can see from table 3 estimator d^D, which takes into account auxiliary information, is not in all cases superior to the usual estimator *d*. It is superior only for *p*=0, *p*=0,2 and *p*=0,5. For *p*=0,8 and for *p*=1 superior is the usual estimator *d*, which does not use any auxiliary information.

For estimator *d* efficiency of the estimation is the highest, that is *MSE* is the lowest for *p*=1, that is for the same sample on two occasions. For estimator d^D efficiency of the estimation is the highest, that is *MSE* is the lowest for *p*=0, that is for disjoint samples on two occasions.

Now we can compare these results with theoretical results from section 2.

On the basis of theorem1 for this population we have:

- $D^2(d^D) \le D^2(d) \Leftrightarrow p \le 0{,}6099$

On the basis of theorem2 we have:

- min $D^2(d^D)$ is achieved for p=0, because $M = -0{,}0935 < 0$.

From statistical theory we have:

- min $D^2(d)$ is achieved for p=1, because $\rho(Y_2, Y_1) > 0$.

As we can see all simulation results agree with theoretical ones.

From table 4 we can see that gain in the efficiency of the estimator d^D compare to the usual estimator d vary from -29,1% to 27,1%.

3.3 Population 3

According to multivariate normal distribution a population of N=5000 elements was generated. Finite population parameters were as follows:

$$\begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 \\ \overline{X}_1 \\ \overline{X}_2 \end{bmatrix} = \begin{bmatrix} 3{,}998 \\ 4{,}990 \\ 2{,}994 \\ 3{,}986 \end{bmatrix}, \quad S = \begin{bmatrix} 0{,}798 \\ 0{,}996 \\ 0{,}600 \\ 0{,}804 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0{,}800 & 0{,}605 & 0{,}402 \\ 0{,}800 & 1 & 0{,}309 & 0{,}505 \\ 0{,}605 & 0{,}309 & 1 & 0{,}704 \\ 0{,}402 & 0{,}505 & 0{,}704 & 1 \end{bmatrix}$$

On two occasions samples of 100 elements were selected according to SRSWR; five different matched fractions of two samples were taken into account: p=0, p=0,2, p=0,5, p=0,8, p=1. For each case sampling was repeated h=1000 times.

Results of a simulation study for population 2 are presented in Table 5 and table 6.

Table 3. Mean Square Error

	p=0	p=0,20	p=0,50	p=0,8	p=1
$MSE(d)$	0,0170	0,0143	0,0103	0,0060	0,0035
$MSE(d^D)$	0,0116	0,0099	0,0067	0,0035	0,0015
$EFFP(d^D)$	46,7%	44,5%	54,2%	71,7%	127,3%

As we can see from table 5 estimator d^D, which takes into account auxiliary information is in all cases superior to the usual estimator d, i.e. $MSE(d^D)<MSE(d)$ for all p. For estimator d efficiency of the estimation is the highest, that is MSE is the lowest for p=1, that is for the same sample on two occasions. For estimator d^D efficiency of the estimation is the highest, that is MSE is the lowest also for p=1, that is for the same samples on two occasions.

Now we can compare these results with theoretical results from section 2.

On the basis of theorem1 for this population we have:

- $\forall p \in <0,1> \quad D^2(d^D) \le D^2(d)$

On the basis of theorem2 we have:

- min $D^2(d^D)$ is achieved for p=1, because $M = 0{,}625 > 0$.

From statistical theory we have:

- min $D^2(d)$ is achieved for p=1, because $\rho(Y_2, Y_1) > 0$.

As we can see all simulation results agree with theoretical ones.

From table 6 we can see that gain in the efficiency of the estimator d^D compare to the usual estimator d vary from 46,7% to 127,3%.

4. CONCLUSIONS

The use of auxiliary information in the form of composite estimators on each successive occasion in repeated rotating surveys and consequently increase of the efficiency of mean estimation on each occasion can cause two different effects. It can cause both increase or decrease of the efficiency of the estimation of net changes, which are commonly estimated together with the population mean in multipurpose rotating surveys. The type of the effect, increase or decrease of the efficiency of net changes, depends on number of matched units and various population parameters. Exact mathematical formulas are given in theorem 2. The second important conclusion that arises from the paper is the following. It is well known that when estimating net changes by using usual estimator, that is difference of two sample means, it is best to retain the same sample for the next occasion when $\rho(Y_{t+1}, Y_t) > 0$. However this well-known theorem is valid only when usual sample means are applied on both occasions. When using composite estimator it is possible that $\rho(Y_{t+1}, Y_t) > 0$ and nevertheless it is best to draw a new sample for the next occasion. In the case of difference estimators and auxiliary information optimum number of matched units maximizing precision of the estimation depends on the combination of all correlation coefficients in the form of M given by (7) and not simply on the correlation coefficient between study variable on two occasions. Simulation studies illustrating above problems are presented in section 3. From all simulation studies it implies that efficiency of the estimation to be gained or lost is considerable; in presented simulation studies it varies from -29,1% to 127,3% compare to usual estimation. Thus importance of presented problems is not only theoretical but also practical.

REFERENCES

Berger, Y.G., Priam, R. (2010). Estimation of Correlations between Cross-Sectional Estimates from Repeated Surveys - an Application to The Variance of Change. Proceedings of The 2010 Statistic Canada Symposium

Binder, D.A. and Dick, J.P. (1989). Modelling and Estimation for Repeated Surveys. *Survey Methodology*, 15 (1), 29-45.

Kim, K. W., Park. Y. S. and Kim, N. Y. (2005). l-Step Generalized Composite Estimator under 3-Way Balanced Rotation Design. *Journal of the Korean Statistical Society*, 34, 219 - 233.

Duncan, G.J. and Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review* 55 (1), 97-117.

Jones, R. (1980). Best Linear Unbiased Estimators for Repeated Surveys. *Journal of the Royal Statistical society Series B* 42, 221 – 226.

Kordos, J. (2012). Review of Application of Rotation Methods in Sample Surveys in Poland. *Statistics in Transition, New Series*, 2012, 13 (2), 47—64.

Kowalczyk, B. (2003). Estimation of The Population Total on The Current Occasion under Second Stage Unit Rotation Pattern. *Statistics in Transition*, 6 (4), 503-513.

Kowalczyk, B. (2003). Classical and Regression Estimator of The Ratio on Successive Occasions, Survey Sampling In Economic and Social Research, University of Economics in Katowice (in Polish), *Katowice*, 78-79.

Kowaczyk, B. (2004). Number Index Estimation using Auxiliary Information in Repeated Rotating Surveys, Sampling Designs for Environmental, Economic and Social Surveys: Theoretical and Practical Perspectives, Conference Papers, September 23-24, Sienna, Italy.

Kowalczyk, B. (2013). Composite Estimation Issues in Sample Surveys Based on Rotating Samples (in Polish), Warsaw School of Economics Publishing.

Kowalski (2009). Optimal Estimation in Rotation Patterns. *Journal Statistical Planning and Inference*: 139(4), 2429-2436.

McLaren, C. H. and Steel, D. G. (2001). Rotation Patterns and Trend Estimation for Repeated Surveys using Rotation Group Estimates. *Statistica Neerlandica*, 55, 221-238.

Park, Y. S., Kim, K. W. and Choi, J. W. (2001). One-Level Rotation Design Balanced on Time in Monthly Sample and in Rotation Group. *Journal of the American Statistical Association*, 96, 1483-1496

Patterson, H.D. (1950). Sampling on Successive Occasions with Partial Replacement of Units. *Journal of the Royal Statistical Society*, *Series B* 12, 241-255.

Rao, J. and Graham, J. (1964), Rotation Designs for Sampling on Repeated Occasions. *Journal of the American Statistical Association* 50, 492-509.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Wesołowski, J., (2010). Recursive Optimal Estimation in Szarkowski Rotation Scheme. *Statistical in Transition*. 11(2) , 267-285.