Estimation of standard error of the parameter of change using simulations

Djordje PETKOVIC
Statistical Office of the Republic of Serbia

ABSTRACT

The main objective of this paper is to present the procedure for estimating standard error of parameter of change (index) of turnover in R software (R core team, 2014) when samples are coordinated. The problem of estimating standard error is dealt with in the statistical literature by various types of approximations. In my paper I start from the method presented at the Consultation on Survey Methodology between Statistics Sweden and Statistical Office of the Republic of Serbia (SERSTAT 2013:22), make simulations and calculate estimate of the correlation and true value of standard error of change between turnovers from two years. I use two consecutive sampling frames of quarterly Structural Business Survey (SBS). These frames are updated with turnover from corresponding balance sheets. Important assumption is that annual turnover is highly correlated with quarterly turnover and that computed correlation can be referred to when comparing methods of estimation of correlation on the sample data.

Keywords: *coordinated samples with permanent random numbers, standard error of index, correlation of totals, simulations, R, quarterly SBS.*

INTRODUCTION

The purpose of this paper is to present a method based on simulations for estimating the standard error of a parameter of change (index) when samples are coordinated by permanent random numbers.

The problem of estimating standard error of an index when samples are coordinated is an open theoretical and practical problem. Coordinated samples are not independent, but overlapping, so this feature has to be taken into account in the estimation procedure.

Work on estimation of standard errors of indices of coordinated samples in the Statistical Office of the Republic of Serbia (SORS), is based on simulations conducted by A. Lindblom and S. Berg within Activity on Survey Methodology of the Project Partnership in Statistics funded by SIDA (Swedish International Development Cooperation Agency). Different methods were compared and evaluated. Some of these methods are based on the estimate of the correlation on the common part of the samples, which overestimates the true value of correlation so that the variance is underestimated. Hence, we used some correction factors.

This paper presents a way in which the simulations can be carried out in R programming language in order to calculate values of correlation between the

estimates of totals and "true" variance of the index. For this purpose I use data from quarterly SBS.

In the first part of this paper I briefly present the frame, sample design and estimation procedure of the quarterly SBS. Then I explain the method for estimation and give details and descriptions of some of the codes I use for simulations.

FRAME, SAMPLE DESIGN AND ESTIMATION PROCEDURE OF SBS

Quarterly SBS is a survey conducted with the aim of examining quarterly dynamics of enterprises' financial data and changes in the economic structure of the non-financial business sector. The population for this survey includes all enterprises with the main activity in the manufacture and sale of goods and services for the market. (Zamaklar et al, 2014). Sample is stratified simple random. It is selected from the sampling frame by a sequential scheme using permanent random numbers. Frames' units are randomly assigned to 5 rotation groups. Rotation is achieved by an annual shift of permanent random numbers of one rotation group for 0.1[1].

Based on the survey data, indices are computed, both annual index (current quarter to the corresponding quarter of the previous year) and the chain index (current to the previous quarter). Some of the outputs of the survey are annual indices by domains defined by sections of NACE Rev. 2. The details are in the Table 1.

Table 1[2]

Domain umber	Domains
1	Total
2	Agriculture, Hunting, Forestry and Fishing
3	Industry and construction
4	Services
5	Wholesale and retail trade
6	Transport and storage
7	Accommodation and food service activities
8	Information and communication
9	Other services

For these simulations I use two data frames, one from 2012 and the other one from 2013. These frames were updated with auxiliary information on turnover from annual financial report of the corresponding years. After updating the frames, turnover values were missing for some units. I did imputations on the level of stratum by giving value of an average of turnover in that stratum plus a disturbance term so that frames used in simulation had no missing values on turnover.

In the simulations, I generate samples that correspond to the real situation in SBS and compute the index of the estimate of turnover for 2013 to the previous year. Based on a large number of simulations (10 thousands), I calculate the correlation between totals and an approximation of the "true" variance of the index of turnover.

1. Sample coordination of statistical business surveys has been introduced in SORS in 2013 (Trpinac Melovski, O. et al (2014) and it is based on the Swedish system SAMU (2013).

2. The categorization of economic activity in this table is based on NACE Rev. 2 and the source is (Zamklar, 2014)

Estimation of the Horvitz-Thompson totals is produced using survey package. I did estimation on the domains level. Formulas for estimation of the variance of indices are presented in the next part.

Because the annual data on turnover, from the final financial accounts, is in the correlation with the quarterly survey data, computed correlation can to be used in the estimation of the variance of index in "real life" situations when only survey sample data is available (see formula (7)).

METHOD

One of the methods that is used in SORS for estimating the variance of index is based on the Taylor linearization. As shown in Verma (1993), variance of the ratio $\hat{r} = \frac{\hat{t}_1}{\hat{t}_0}$, where $\hat{t}_0$ is the level estimate for year 0 and $\hat{t}_1$ is the level estimate for year 1 can be expressed as

$$V\left(\frac{\hat{t}_1}{\hat{t}_0}\right) \approx V(\hat{r}) \approx \frac{1}{(\hat{t}_0)^2} \cdot \left(V(\hat{t}_1) + \hat{r}^2 \cdot V(\hat{t}_0) - 2 \cdot \hat{r} \cdot Cov\ (\hat{t}_0, \hat{t}_1)\right) \tag{1}$$

Using the relationship between correlation and covariance

$$\rho(\hat{t}_0, \hat{t}_1) = \frac{C(\hat{t}_0, \hat{t}_1)}{\sqrt{(V(\hat{t}_0) \cdot V(\hat{t}_1))}} \tag{2}$$

we can rewrite the formula:

$$V\left(\frac{\hat{t}_1}{\hat{t}_0}\right) \approx \left(\frac{\hat{t}_1}{\hat{t}_0}\right)^2 \cdot \left(\frac{V(\hat{t}_1)}{\hat{t}_1^2} + \frac{V(\hat{t}_0)}{\hat{t}_0^2} - 2\frac{\rho(\hat{t}_0, \hat{t}_1)\sqrt{(V(\hat{t}_0) \cdot V(\hat{t}_1))}}{\hat{t}_0 \hat{t}_1}\right) \tag{3}$$

We can see from formula (3) that the estimation of variance is reduced to the calculation of variance of totals and correlation between totals in two different periods. In real life these values are estimated, but by large number of simulations we can compute the "true" values.

Formulas that are used for calculation of variance and covariance from simulations are:

$$V(\hat{t}_0) \approx \frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{t}_{0k} - \bar{\hat{t}}_0\right)^2 \tag{4}$$

$$V(\hat{t}_0) \approx \frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{t}_{0k} - \bar{\hat{t}}_0\right)^2 \tag{5}$$

$$V(\hat{t}_1) \approx \frac{1}{N-1}\sum_{k=1}^{N}\left(\hat{t}_{1k} - \bar{\hat{t}}_1\right)^2 \tag{6}$$

These formulas are the ones that I used in R simulations. Even though this procedure can't be entirely applied in "real life" situations, there are parts of it that

can be used. Most importantly, correlation calculated with simulations according to the formulas (2), (4) - (6) can be re-used in formula

$$\hat{V}\left(\frac{\hat{t}_1}{\hat{t}_0}\right) \approx \left(\frac{\hat{t}_1}{\hat{t}_0}\right)^2 \cdot \left(\frac{\hat{V}(\hat{t}_1)}{\hat{t}_1^2} + \frac{\hat{V}(\hat{t}_0)}{\hat{t}_0^2} - 2\frac{\rho(\hat{t}_0,\hat{t}_1)\sqrt{\left(\hat{V}(\hat{t}_0)\cdot\hat{V}(\hat{t}_1)\right)}}{\hat{t}_0\hat{t}_1}\right) \quad (7)$$

Where $\hat{t}_0$, $\hat{t}_1$, $\hat{V}(\hat{t}_0)$ and $\hat{V}(\hat{t}_1)$ are estimates based on a specific "real life" sample.

COMMENTS ON THE R SIMULATIONS

In the codes below I present the selection of the sequential random samples based on permanent random numbers from the uniform distribution. I put the names of the variables in the brackets.

Firstly, I do imputations for both of the frames, f1 and f2. I use `ddply()` function to calculate the mean (mean) and standard deviation (sd) in each stratum (strat) and put the results in data set t1. After it, I merge frame f1 and frame t1 in order to impute those values of turnover (turnover1) that are NA. I imputed $mean + sd * {unimp}/{4}$, $unimp : N(0,1)$ generated with `rnorm`. I did imputations for the second frame in the same way.

```
> t1 <-
ddply(f1,~strat,summarise,mean=mean(turnover1,na.
rm=TRUE),sd=sd(turnover1,na.rm=TRUE))
> f1 <- merge(t1,f1,by="strat")
> f1$unimp <- rnorm(length(f1$id), mean = 0, sd = 1)/4
> f1$turnover1[is.na(f1$turnover1)] -> f1$mean[is.
na(f1$turnover1)]+f1$sd[is.na(f1$turnover1)]*f1$unimp[is.
na(f1$turnover1)]
```

The first frame, f1, is filled with random numbers from uniform (0,1) distribution. Each unit with its unique id gets random number (prn). After it, the frame is ordered in ascending order according to these permanent random numbers inside each strata using with() function. In each strata, first pnh1 units is sampled (pnh1 is sample size selected from each stratum h of size Nh1). Samp1 is the selected sample. The code from R is:

```
> f1$prn <- runif(length(f1$id),0,1)
> f1 <- f1[order(f1$strat,f1$prn),]
> f1$br  <- with(f1, ave(f1$strat,f1$strat, FUN = seq_along))
> samp1 <- f1[as.numeric(f1$br)<=f1$pnh1,]
```

After the first sample is selected, the frames from both years are merged (up12u) with `merge()` and those units from the second frame that are in the first as

well get the same permanent random number they had in the first frame. After that the rotation is performed on those units from intersection that are in the first rotation group. Those units that don't have PRN after merging get a new one and the selection of the second sample is done in the same way as the selection for the first sample. Function `sapply()` is used to identify those units from the second frame that are not in the intersection.

This is code for rotations:

```
> up12u$cif <- as.numeric(substring(as.character(1000000000*up12u$prn)
,6,7)) %% 10
> up12u$rgr <- up12u$cif %% 5
> up12u$prn[up12u$prn+0.1 < 1 && up12u$rgr==1] <- up12u$prn+0.1
> up12u$prn[up12u$prn+0.1 < 1 && up12u$rgr==1] <- up12u$prn+0.1-1
```

This is the selection of the second sample based on coordinated samples:

```
> up12u <- merge(f1,f2,by="id", all=TRUE)
> up12u.prnna <- up12u$id2[is.na(up12u$prn)]
> index <- sapply(up12u.prnna, function(x) which(up12u$id2 == x))
> prn <- runif(length(up12u1$prn),0,1)
> up12u$prn[is.na(up12u$prn)] <- prn[index]
> f2_pom <- up12u[!is.na(up12u$ind2), ]
> f2 <- f2_pom[order(f2_pom$strat.y,f2_pom$prn),]
> f2$br2 <- with(f2, ave(f2$strat.y, f2$strat.y, FUN = seq_along))
> samp2 <- as.data.frame(f2[as.numeric(f2$br2)<=f2$pnh2, ])
```

After I select samples, I make nine domains based on Table 1. I use the function subset() for defining 9 subsets from the first sample. Variables saopred1 and saopred2 define domains. This is the code for the first sample:

```
> ds11 <- samp1
> ds21 <- subset(samp1,saopred1==2)
> ds31 <- subset(samp1,saopred1==3)
> ds41 <- subset(samp1,saopred1==5 | saopred1==6 | saopred1==7 |
saopred1==8 | saopred1==9)
> ds51 <- subset(samp1,saopred1==5)
> ds61 <- subset(samp1,saopred1==6)
> ds71 <- subset(samp1,saopred1==7)
> ds81 <- subset(samp1,saopred1==8)
> ds91 <- subset(samp1,saopred1==9)
```

The sub setting is done in the same way for the second sample, samp2.

I get estimates of the totals using survey package (Lumley, 2014) (svytotal() function), after I define the design with svydesign().Here I give an example for the definition of the survey design for the first domain in the first sample and estimation of the totals of turnover.

```
> ddom11 <- svydesign(id=~1, strata=~strat,data=ds11,fpc=~nh1)
> tott11[i] <- svytotal(~ds11$turnover1,ddom11)[1]
```

I store all the totals from all simulations in one data frame (tts). I repeat the procedure of selecting the samples and estimating the totals for 10000 times. In this way I "mimic" the real life situation 10000 times in order to get estimates. In the end I calculate the variances and correlations based on formulas (2), (4) – (6) and get the results using the final formula (3) for the variance of indices.

```
> tts$var1 <- (tts$tott12/tts$tott11)^2*(tts$v11/tts$tott11^2+tts$v12/
tts$tott12^2-2*tts$corr1*sqrt(tts$v11*tts$v12)/(tts$tott11*tts$tott12))
```

tott11 and tott12 are the estimates of the totals from the first and second sample respectively in the first domain, while v11 and v12 are estimates of variance. Corr1 is the estimate of correlation between totals in two time periods.

CONCLUSION

In this paper I presented a way of calculating approximation of the true correlation and variance between totals in two consecutive time periods when samples are coordinated by permanent random numbers. I used simulations for this purpose. I implemented the method for estimation in R, using survey package. Based on simulations I computed the correlation between totals and the "true" variance of the index of turnover. Aside from the benefit of having programs for estimating the variance of index, this project produced values for correlation that can be used for further estimation when only survey data is available for quarterly SBS.

Computation of the "true" values is just part of the activity concerning the estimation of the variance of index in SORS. Various other methods are investigated by simulations and compared between themselves. The values produced in the simulations that are presented in this paper are benchmarks for other methods. One of the future projects would be to implement some of the methods for estimation of variance of change parameter in R as well.

REFERENCES

1. Melovski Trpinac, O., M.Ninić, M.Panović (2014) Sampling Coordination of Statistical Business Surveys, Working paper of the Statistical Office of the Republic of Serbia, No. 89, Year L (2014).
2. Mission reports of the Statistics Sweden International Consulting Office: SERSTAT 2010:09, SERSTAT 2013:22
3. Ohlsson, E. SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden. R&D Report, Statistics Sweden, 18 (1992).
4. R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austira. URL http://www.R-project.org/.
5. SAMU – The system for co-ordination of frame populations and samples from the Business Register in Sweden Department of Economic Statistics, Statistics Sweden (2003).
6. T. Lumley (2014) "survey: analysis of complex survey samples". R package version 3.30.
7. Verma, V. Sampling errors in household surveys, UN (1993).
8. Zamaklar, G., O. Melovski Trpinac, M. Quarterly business activities of enterprises in the Republic of Serbia, Working paper of Statistical Office of Republic of Serbia, No. 88, Year L, (2014).