Standard error estimation – how to do it quickly, efficiently and correctly

Rudi Seljak[1], Jerneja Pikelj[2], Petra Blažič[3]

[1]*Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia, rudi.seljak@gov.si*
[2]*Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia, jerneja.pikelj@gov.si*
[3]*Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia, petra.blazic@gov.si*

Abstract

Standard error estimation is in the case of sample surveys one of the most demanding and challenging tasks during the survey evaluation process. On the other side it is also true that in sample surveys standard error can still be considered as a key quality indicator that predominantly indicates the accuracy of the disseminated results. It is therefore of crucial importance that these indicators are on the disposal for the estimates, they are on disposal quickly and are based on the sound and transparent methodology. It is a special challenge, how to fulfil all these requests in the situation when we are more and more facing demands for more results, shorter timeliness and lower costs.

At the Statistical Office of the Republic of Slovenia we spent last few years to develop generic software for standard error estimation. The developed metadata driven application should enable that the standard errors are on disposal for all the statistical results that are based on the data from random sample and that they are on disposal at the same time as the results themselves. During development of such general tool, which aims to cover a wide range of different input data and different estimators, we faced many methodological and practical challenges. Many times pragmatic solutions to the problems outweighed the more "theoretically clean" ones. In the paper we present the main principles of the general application, describe the main challenges in its development and how we dealt with them.

Keywords: standard error, estimators, metadata driven application

1. Introduction

Although in the recent years in the field of official statistics statistical surveys have more and more and more diverse means for data collection (especially from different secondary sources), the "classical" sample surveys are still a very frequently used means for the collection of the needed data. One of the consequences of the use of random samples as the set of observational units is that all the results derived from such a survey contain the sampling error, usually expressed as a standard error. It is a fact that, in the modern quality assessment models the precision of the statistical results is by no more the prevailing (or even the only) quality dimension, but in fact it still remains a very important criterion of the quality of the statistical results. Therefore the national statistical organisations have an important obligation to correctly and accurately estimate the standard errors whenever the sampling approach is used and then to present these errors to the data users in the transparent and clearly readable form.

In the case of sample surveys, standard error estimation is still a very demanding task and it presents a big challenge from the theoretical as well as from the practical point of view. Many different approaches and practices have been developed in the last decades. Many institutions and individuals were facing these problems and then contributed to this quickly developing area of the inferential statistics. Parallel with the theoretical development in this area, a wide range of computer applications, which implement the before mentioned methods, were designed and developed. So, these days the statisticians have several practical tools at their disposal when they deal with variance estimation problems. But despite this fact, there is still a big challenge, how to use these tools in a systematic and automated way for all the different estimators and sampling design that can take place in the field of official statistics.

Besides the estimation of the standard error, another challenge is how to present these errors to the users. The first goal is to achieve that the practice is standardised, meaning that it is done by the same system for all the surveys. The second goal is to achieve that the errors are presented in transparent and clear way, meaning that they are understood by a wide range of different users of our results.

At the Statistical Office of the Republic of Slovenia (hereinafter SURS) the problem of estimation and dissemination of standard errors has been challenged for many years. In the past the calculation of the sampling errors was quite »survey dependent«. Each survey had its own system, which was mostly dependent on the survey methodologist and there were no general rules to be followed. Usually the direct estimations of the sampling errors were performed only for the key statistics and for the key domains, while for the other statistics and (sub) domains some simple (linear) models were used. Also the standard errors for the estimated statistical results were also very rarely published explicitly, only the results with lower degree of precision were marked and the coefficient of variation was the “exclusive” criterion, used for the “marking” system.

To overcome the survey dependent and non-standardized practice, a few years ago a significant revision of the system was carried out. In the framework of this revision, the following steps were taken:

- General rules were set up for the sampling error estimation for the different types of estimators as well as for the different sampling designs. To enable the standard system based on the general rules, a certain degree of simplification had to be employed for some procedures.
- New rules were set up for the dissemination and presentation of the sampling errors.
- A special general application was developed in which all the above mentioned rules were incorporated. The main goal of the application was to enable quick, efficient and unified sampling error estimation.

In the paper we focus on the general application, presenting its main principles and describing the main challenges in its development. In the following chapter we briefly describe the general concept of the metadata driven application that is also used for the implementation of other phases of the statistical process. In the last part we present the most outstanding challenges that we faced in the course of implementation of the generic software application and pragmatic solutions used to deal with them.

2. General application

The general application was developed at SURS in order to support different steps, of the statistical production, ranging from data validation and data editing to data aggregation and tabulation. The whole system is based on a set of small generic solutions, which are designed in a way that they enable easy and flexible linking of inputs and outputs of the individual components to the whole statistical process. These components, which we also call the building blocks, provide the generic software solution for the certain part of the statistical chain and are designed in a such way that they can act independently from each other. The main features of these building blocks could be summarized as follows:

- They are designed on the basis of harmonized, transparent and widely accepted methodological principles, which have been determined before the actual creation of the particular building block.
- The building blocks can be plugged to different databases in different environments (e.g. ORACLE, SAS) as long as the databases follow some basic rules for the organization of the data.
- They are designed as fully metadata driven systems, meaning that information which determines the parameters for the execution of the processing for the concrete survey and concrete reference period are provided outside the core computer code.
- There is one single, unique database of process metadata. This database is created in ORACLE and managed by the .NET application, which enables user friendly management of the process metadata.

When the data aggregation phase of the process is concerned, another important feature of the general application has to be pointed out. Namely, the fact that the application “merges” several processes (aggregation, sampling error estimation, disclosure control, tabulation) that have previously been conducted separately, into one fully automated process. In the following chapters we present two of the many challenges that we faced in the stage of the implementation of the general application for the case of the particular surveys.

3. Challenges associated with defining metadata for domains and statistics

In the implementation of statistical data processing we want to estimate the value of the unknown population parameters, called statistics, and the associated precisions. We are often interested in the value of statistics, not only for the whole population, but also for its subsets, called domains. The general application, presented above, enables the calculation of various types of statistics - such as totals, means, percentiles, ratios for different domains. Besides the calculation of statistics, the application also provides the estimation of their standard errors. The calculations are carried out by using the SAS "proc surveymeans" procedure.

The application currently enables the calculation of the following types of statistics:

Type of Statistic	Description
01	Number of units with a certain characteristic
02	Proportion of units with a certain characteristic
03	Total
04	Mean
05	Ratio
06	Percentile

Table 1: Types of Statistics

In order to show the main issues, related to the definition of the process metadata for aggregation, we will look at the example on the data of the Community Innovation Survey (CIS). Suppose that we want to estimate the number of innovative enterprises by size classes. We can define the metadata for the statistics and the domains in two different ways:

1. Statistics is calculated by using the dummy variable named INOV, defined as:

$$\text{INOV}=\begin{cases}1;\ \text{if the enterprise is innovative}\\ 0;\ \text{otherwise}\end{cases} \qquad (1)$$

The domain is defined with the categorical variable SIZE_CLASS, determined by the number of employees:

$$\text{SIZE\_CLASS}=\begin{cases}1;\ \text{if the number of employees is greater than 250}\\ 2;\ \text{if the number of employees is between 50 and 250}\\ 3;\ \text{if the number of employees is less than 50}\end{cases} \qquad (2)$$

2. In the second case, the statistics is calculated on the variable named ONE, where all units have value one.

$$\text{ONE}=1 \qquad (3)$$

The domain is defined as the combination of the variables INOV defined in (1) and SIZE_CLASS defined in (2).

First, we have to define the derived variables, described above. The application also provides the possibility of derivation of new variables. The process metadata that define the derived variables are given in the table below:

Variable_label	Condition	Value
SIZE_CLASS	If employee>=250	1
SIZE_CLASS	If 50<=employee<250	2
SIZE_CLASS	If employee<50	3
INOV	If INNOV_ACTIVE in (1, 2)	1
INOV	If INNOV_ACTIVE not in (1, 2)	0
ONE	If 1=1	1

Table 2: Process metadata for the derived variables

Short description of the fields:

Variable_label Label of the derived variable.
Condition Condition which determines for which units a certain rule will be applied.
Value Value of the derived variable.

Now we can finally define domains and statistics. We will denote STAT1 and DOM1 statistics and domain for the first example and STAT2 and DOM2 statistics and domain for the second example. As it can be clearly seen from the definition of domains, the first domain is one dimensional and the second domain is two dimensional.

The process metadata that define domains are given in the table below:

Domain_label	Dom_var1	Dom_var2
DOM1	SIZE_CLASS	
DOM2	SIZE_CLASS	INNOV

Table 3: Process metadata for the domains

Short description of the fields:

Domain_label Label of the domain.
Dom_var1, Dom_var2 List of the variables which define the dimensions of the domain.

The process metadata that define statistics are given in the table below:

Stat_label	Variable	Type
STAT1	INNOV	01
STAT2	ONE	01

Table 4: Process metadata for statistics

Short description of the fields:

Stat_label	Label of the statistics.
Variable	Name of the variable, needed for the calculation of the statistics.
Type	Type of statistics according to a standard code list described above.

In both cases we get the same estimate of the number of innovative enterprises, but they differ in estimated standard errors and number of units on the basis of which the estimated values were calculated. It is also very important which type of statistics we use. We could use type 03 and get the same value of statistics, but the associated precision would be slightly different. This occurs because we assume that we know the population size, when we calculate the number of units with a certain characteristic, and we take the standard error as the criteria for precision. When we calculate the population total (type of statistics 03), we do not use the same assumption, and take the coefficient of variation as the criteria for precision.

Therefore, it is important to we pay attention to the way we define the metadata and to the selection of the type of statistics because the calculation of precision takes into account different procedures and assumptions. In the case above, we decided for the first option, assuming the hypothesis of a pre-known population size and the criterion for precision with respect to the standard error.

4. Challenges associated with the form of the microdata database

The application provides the estimated values of unknown population parameters and estimation of their precisions by using the SAS proc surveymeans procedure. Therefore procedure the structure of input microdata must follow certain rules. The most important rule is that the microdata are organized in standard format, where all the data of one unit are written as one record.

In some cases it is much easier to organize the microdata-database in the transposed form. This means that the data for the responding units is written as several records in the database. We will call this database, the database with holdings, where holding is the characteristics, that uniquely identifies a record within the responding unit.

Let's have a look at an example of the microdata from the field of agriculture statistics. We can organize our microdata in two different ways. The first one is the standardly organized database, where one record represents one observed unit. The first variable is the “id variable” and all the remaining variables tell us the area for a specific culture. The second table represents the database with holdings, and as it is clearly seen, the id variable is duplicated. The second variable identifies the culture and the last variable provides data on the area.

a) Standardly organized database

id	area_culture_11	area_culture_12	area_culture_21
1	A	B	C
2	D	E	.
3	F	G	H

Table 5: Standardly organized database

b) Database with holdings

id	culture	area
1	11	A
1	12	B
1	21	C
2	11	D
2	12	E
3	11	F
3	12	G
3	21	H

Table 6: Database with holdings

Microdata organized in such a way in the database are not appropriate to carry out the aggregation with our application. In this case the application would give us wrongly estimated values for population means, proportions, quantiles, their precisions and the number of units on the basis of which the estimated values were calculated. Only the population total would be estimated correctly.

Use of the database with holdings (b) together with the existing general programs does not give us the right calculations, but such a generalization of the structure of the input microdata, would be a great benefit for the users of the application. Namely, in this case the defining of the metadata is much easier and faster.

What follows is a short illustrative example of the usage of the process metadata for calculation of population totals of the areas by different cultures in two different ways, depending on the structure of the input microdata.

a) Database in standard format:

When the database of microdata has standard form, the number of statistics is the same as the number of different cultures. In our case we have three different cultures, meaning that we have to define three different statistics:

Stat_label	Variable	Type
Label_area11	area_culture_11	03
Label_area12	area_culture_12	03
Label_area21	area_culture_21	03

Table 7: Process metadata for statistics

b) Database with holdings:

In the second case, the microdata-database does not have the standard form. But in this case it is much easier to define the process metadata. We can define only one statistics and one domain. Statistics is defined with the variable area and domain with the categorical variable culture.

Stat_label	Variable	Type
Label_area	area	03

Table 8: Process metadata for statistics

Domain_label	Dom_var1	Dom_var2
Label_culture	culture	

Table 9: Process metadata for domains

In general, when our categorical variable has numerous categories the transformation of the microdata from the form with holdings to the standard form, would result in a new microdata table with several variables. This means, that we would have to define a lot of new statistics. If we find the solution where we can keep nonstandard form of the microdata database, we would achieve great benefit in order to define the process metadata.

The idea is to find a solution that would not require transposing the database back to a standard form (a) and would also provide the correctly calculated value of the statistics, the associated precision, the number of units on the basis of which the calculation was performed, and the weighted number of units.

We have got precisely this solution, by using the assumption of two-stage sampling design. If we pretend that our sampling design was two-stage and that we have chosen a specific unit of observation (defined with ID) at the first stage, and all the associated records for the unit at the second stage, then our calculations will be correct. Thus, we have a solution that gives us properly calculated statistics and their standard errors and also allows the definition of metadata for aggregation to be easy and fast.

Conclusions

Development of a general tool and its introduction into the statistical production certainly imposes many changes in the implementation of the statistical production at the very general level. Successful adjustment of the implementation of the statistical surveys as well as successful adjustment of the organization and functioning of the whole institution is a clear challenge for SURS in the following years.

5. References

L. Lyberg et al. (1997), Survey measurement and Survey Quality, Wiley, N.Y.

Sarndal Carl-Erik (1992), Model Assisted Survey Sampling, Springer-Verlag, N.Y.

Seljak, R., Blazic, P. (2011), Sampling error estimation – SORS practice, paper presented at the 2nd European Establishment Statistics Workshop, Neuchatel, Switzerland

Seljak R. (2014), Metadata driven application for data processing – from local toward global solution, paper presented at the UNECE Work Session on Statistical Data Editing, Paris, France