Sampling Error Estimation – SORS practice

Rudi Seljak, Petra Blažič
Statistical Office of the Republic of Slovenia

1. Introduction

Assessment of the quality in the official statistics has faced significant development in the recent years. If once the quality of the statistical results was mostly assessed through the criteria of accuracy, it is by the modern approaches considered as a multi-dimensional concept, where accuracy is just one of the dimensions. The model which is widely accepted inside the European Statistical System (ESS) hence defines six quality dimensions: Relevance, Accuracy, Timeliness and Punctuality, Accessibility and Clarity, Comparability, Coherence. The important part of the quality assessment is the calculation of the quality indicators for each of the above listed dimensions. At SORS we defined the list of standard quality indicators which should be calculated (if relevant) for each of the survey. The key values of these indicators are reported in the metadata provided together with the disseminated results and the whole list of values is reported in the standard quality report. Some of the standard quality indicators have a special role since they could be derived directly from the statistical process (e.g. editing rate, non-response rate). We call these indicators process quality indicators and we try to incorporate their calculation into the statistical process, so they can be available at the same time as the statistical results. In such way, the indicators can serve as a strong tool for monitoring and controlling of the process.

Although the sampling error is now just one of the many quality indicators, it is for the most of the sampling surveys still the most indicative quality information. Hence it is duty and obligation of the producers of the official statistics that these errors are correctly and timely estimated and presented to the users. Strictly speaking should sampling error be considered more as a product oriented quality indicator, but at SORS it is considered as one of the key process indicators. This is mainly due to the fact that we try to incorporate its calculation directly into the statistical process.

In the past the calculation of the sampling error 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 very rarely explicitly published, only the results with lower degree of precision were marked and the coefficient of variation was the “exclusive” criteria, 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:

- The 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, certain degree of simplification had to be employed for some procedures.
- The new rules were set up for the dissemination and presentation of the sampling errors.
- A special (sas) application was built 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.

Since the theoretical rules for the estimation and dissemination of the sampling errors were already presented in some previous papers (see Seljak, 2008), this papers focuses on the presentation of the IT application. Hence in the next section the application will be described and in the last part some plans for the future improvements will be presented.

2. Application for aggregation, sampling error estimation and tabulation

When we planned the development of the application our main goal was to create a tool where all the above mentioned theoretical rules would be implemented, the whole procedure would be fully automated and its outputs would enable easy and user friendly preparation of the different dissemination outputs. The main features of the application are:

- The application "merges" the processes of aggregation, sampling error estimation and tabulation into one fully automated process.
- The application is designed as a metadata driven (MDD) system, meaning that all the information which determine the parameters for the execution of the processing for the concrete survey and concrete reference period are provided outside the core computer code. No information referring to concrete survey execution is incorporated in the program code but is provided through the special metadata tables.
- After the final phase of the development the application should be manageable only by the subject-matter personnel. No additional input from the IT department would be needed during the execution of the process.

As we mentioned before, the main characteristics of the application is that it is based on a metadata driven principle. The application hence consists of the following three main elements:

- Metadata tables where the parameterization for the particular survey and for the particular execution of the survey is given. At the moment the metadata are stored in the MS Access databases, but should very soon be transferred into ORACLE database.
- The core (SAS) program code. The code is general, meaning that it should never be changed for the needs of the particular survey. All the (meta)data needed for the execution of the particular survey should be given in the outside environment (metadata tables).
- Outputs in the form of Excel spreadsheets. In the outputs all the needed aggregates are given as well as the following metadata produced during the process: sampling error, coefficient of variation, number of the (sample) units used for the calculation of the estimate, denotation in the case of the estimate with lower degree of precision.

2.1 Process metadata description

There are several types of statistics which could be estimated by the application; here we give the example of the three the most frequently used ones: proportions, totals and ratios. We will describe the required metadata through a hypothetical example. Let us assume that we have a sample survey of enterprises with the following variables:

Emp	Number of employees
Turn	Turnover
Wpage	Does the enterprise has its webpage (yes/no)
Nace2	Nace 2-digit group
SizeC	Size class

Let us further assume that we would like to estimate the following statistics:

STAT01	Proportion of enterprises with its webpage
STAT02	Total turnover in enterprises with its webpage
STAT02	Turnover per employee in enterprises with its webpage

The metadata are provided in several separate tables, but the most important are the following:

- **Description of the statistics**. The main information on the statistics are given.

Table	Stat_code	Stat_desc	Type	Dummy	Variable	Variable_en	Variable_den
Table1	STAT01	Proportion of enterprises with its webpage	02	Dummy01			
Table1	STAT02	Total turnover in enterprises with its webpage	03		Var02		
Table1	STAT03	Turnover per employee in enterprises with its webpage	05			Var02	Var03

Table	SAS table with all the micro-data needed for the estimation procedure
Stat_code	Code of the statistics – must be unique inside one table
Stat_desc	Textual description of the statistics (not required)
Type	Type of statistics according to a standard code list (e.g. 02 – Proportion)
Dummy	Name of the Dummy variable needed for the calculation of the proportion. Dummy variable is a variable with 0,1 values. The variable can be already given in the micro-data table or could be calculated in the process by the rules given in the separate table
Variable	Name of the variable required for the calculation of the total. The variable can be given in the micro-data table or could be calculated in the process by the rules given in the separate table
Variable_en	Name of the variable in the enumerator, required for the calculation of the ratio. The variable can be given in the micro-data table or could be calculated in the process by the rules given in the separate table
Variable_den	Name of the variable in the denominator, required for the calculation of the ratio. The variable can be given in the micro-data table or could be calculated in the process by the rules given in the separate table

- **Description of the dummy and derived variables.**

In this table the rules for the calculation of the variables which are not in the "basic" table but are needed in the estimation process are given.

Table	Var_name	Condition	Value
Table1	Dummy01	If Wpage='yes'	1
Table1	Dummy01	If Wpage='no'	0
Table1	Var02	If Wpage='yes'	Turn
Table1	Var02	If Wpage='no'	0
Table1	Var03	If Wpage='yes'	Emp
Table1	Var03	If Wpage='no'	0

Table	SAS table with all the micro-data needed for the estimation procedure
Var_name	Name of the derived variable
Condition	Condition which determines for which units certain rule will be applied
Value	Value of the derived variable

- **Description of the domain variables**

Usually the statistical estimates are not required just for the whole observed population but also for some sub-populations also called domains. In our case we will require results for one 1-dimensional domain, determined with Nace2 class and one 2-dimensional domain, determined with Nace2 class and Size class. Up to 10-dimension domains can be defined in the application.

Table	Domain_code	Dom_var1	Dom_var2	• • •	Dom_var10
Table1	Dom1	Nace2			
Table1	Dom2	Nace2	SizeC		

Table	SAS table with all the micro-data needed for the estimation procedure
Domain code	Unique code of the domain
Dom_var1-Dom_var10	List of the variables which define the dimensions of the domain.

Beside the above described metadata tables, there are several more tables, needed for the process execution. Here we just briefly explain which information are also needed:

- Information on sample design, strata and primary sampling units (if used)
- Sampling rate for each of the stratum cells
- Type of criteria used for the denotation of the statistics with lower precision
- Form and content of the output tables

2.2 SAS application

The core part of the application is the SAS program where all the processing (according to the provided metadata) is performed. Application consists of three parts (three SAS macros), each one of them has its own standard input and standard output. The three parts are:

- **Calculation of the derived and dummy variables.** The input is the "basic" micro-data table. Output is the input table supplemented with the derived and dummy variables.
- **Aggregation and sampling error estimation**. The input is the output table from the previous process. The output is the table with all the aggregates and referring metadata in the raw form (see next section).
- **Creation and designing of the tables**. The input is the output table from the previous process. The outputs are the Excel spreadsheet tables, designed for the dissemination.

2.3 Output tables

Two types of outputs are provided. The first one is the table with all the aggregates and referring metadata in the "raw" form. Each row of the table gives the information on one aggregate. The following information are given in the table:

DOM1	Variable which determine the first variable in the definition of domain
DOM_VAL1	Value of the first domain variable
…	
DOM10	Variable which determine the tenth variable in the definition of domain
DOM_VAL10	Value of the tenth domain variable
STAT_CODE	Code of the statistics
STAT_VAL	Estimated value of the statistics
NUM_UNIT	Number of units, used for the calculation of the estimate
SE	Estimated sampling error
CV	Coefficient of variation
VAL_DISS	Estimated value, formatted for disseminated (together with the eventual denotation for the estimates with lower precision)

The second group of outputs represent the tables that are to the certain degree already designed for the dissemination purposes. Each table is created in four different versions, providing the following aggregate information: value of statistics, sampling error, coefficient of variation and value of statistics in the form for dissemination (with eventual denotations). The design of the tables is

determined in before described metadata tables. For instance in the case of 2-dimesional domain from our hypothetical case, the four tables would probably be designed as it is presented in the following picture:

Figure 1: Four different table outputs

	Domain variable 2
Domain variable 1	**Estimated aggregates**

	Domain variable 2
Domain variable 1	**Estimated sampling errors**

	Domain variable 2
Domain variable 1	**Estimated coeffcients of variation**

	Domain variable 2
Domain variable 1	**Disseminated aggregates**

3. Conclusions

The application described in the paper represents an important achievement in the process of the modernization of the statistical processes at SORS. The two processes, aggregation and sampling error estimation, which have formerly been separated and managed by ad-hoc procedures, are now merged together in one process, which is based on standardize procedures and fully automated. The additional advantage of the application is the fact that it can be managed only by the subject matter personnel and that represents a significant rationalization of the (human) resources needed for the execution of the survey. However, the application is still in the development phase and some additional improvements will be needed in the future, especially:

- Development of the user interfaces for the easier management of the application, especially for the insertion of the required metadata.
- Development of the expression builder for the arithmetic and logical expressions in order to decrease the current high risk of syntax errors.
- Transfer of metadata database into ORACLE environment.
- Supplementation of the application functionalities with the possibility to estimate the sampling error for indices.

REFERENCES

Kish Lesley: Survey sampling: John Wiley & sons, 1965

Lyberg L. et al.: Survey Measurement and Survey Quality, Wiley, 1997.

Seljak R.: Standard Errors Presentation and Dissemination at the Statistical Office of the Republic of Slovenia; Paper presented at the European Conference on Quality and Methodology in Official Statistics, Rome, Italy, 8-11 July, 2008