Draft version

A tool for sample selection to be used in a standardised production process

Annika Lindblom
Statistics Sweden
SE-701 89 Örebro, Sweden

1. Introduction

The majority of the repeated business surveys at Statistics Sweden (SCB) use the so called SAMU-system for different steps related to creation of frame populations and random samples from the Statistical Business Register (SBR). The absolute majority of the surveys included in the SAMU-system are part of the economic-statistical system with the National Accounts (NA) as a main user. The SAMU-system is restricted to coordinated random samples based on four frozen versions of the SBR each year. Surveys outside the SAMU-system currently use different kinds of tools for sample selection (and the other related steps). SCB is now in the process of developing a general and much more flexible tool for sample selection. A tool possible to use for all sorts of frame populations and samples: coordinated, independent, multistage and multiphase drawn from any kind of register. This paper presents the SAMU-system, experiences and requirements behind the new tool and finally, the new tool for sample selection.

2. The SAMU-system

2.1 The SAMU-system in general

The SAMU-system (see Lindblom 2003) is a VB6-application for different steps related to sample selection focused on surveys within the economical-statistical system. Surveys within the economical-statistical system in particular requires coordination of frame populations and samples because output from these surveys constitute pieces to complete the Gross Domestic Product (GDP). There are three main objectives of using the SAMU-system: (1) to obtain statistics comparable both in time and between surveys; (2) to ensure high precision in estimates of change over time; and (3) to spread the response burden among the businesses.

The SAMU-system is based on four frozen versions of the Statistical Business Register (SBR) each year. Originally based on a frozen version of the SBR but additional information is provided as well as some re-classification of units (based on requirement mainly from the NA) in order to obtain the so called S*tatistical Frame* (SF). The SF is the base for all surveys in the SAMU-system. Currently about 30 surveys use the SAMU-system to create frame populations and select random samples.

2.2 Sampling designs and coordination in the SAMU-system

There are only two sampling designs available in the SAMU-system: coordinated (stratified) simple random samples (see Ohlsson 1992 and 1997) and coordinated (stratified) Pareto πps samples (see Rosén 2000**)**. Samples are drawn by the sequential technique and the basic idea to obtain coordination is to associate a *permanent*, independent and unique random number (PRN), uniformly distributed over the interval (0,1), with every unit in the SF. For every unit persisting in the SF the same random number is used on each

sample occasion. In this way we always get a (stratified) simple random (Pareto πps) sample from the up-dated SF but a large overlap with the latest sample can be expected.

By the symmetry of the uniform distribution the first n units to the left, or to the right, of any fixed point in the interval (0,1) could be selected. The possibility to choose starting point and sampling direction enables coordination between surveys, in the first place used to spread the response burden among the businesses. However, the positive coordination over time for repeated surveys makes it possible that a selected unit may have to participate in a survey for many years. In order to spread the response burden between units (and partly renew the sample) a system for sample rotation has been implemented into the SAMU-system. All units in SF are randomly assigned to one (out of five) rotation groups and each year, in one of the groups, all PRN are shifted.

2.3 Functionality of the SAMU-system

The SAMU-system has the following built in features:

1) Creation of four versions of the SF each year (including assigning and shifting PRNs) based on frozen versions of the SBR
2) Maintenance of survey specific information like survey identification (a unique number), survey name, to which block (starting point and sampling direction) the survey is assigned
3) IT-support for input of the survey design for a specific survey:
   - delimit the frame population from one SF-version
   - stratify this frame population
4) IT-support for guidance when deciding sample sizes (Neyman allocation)
5) IT-support for input of decided sample sizes
6) IT-support for drawing coordinated samples as described in in section 2.2
7) A database for general and unified storage of frame populations and samples for each survey and sample occasion

2.4 Limitations of the SAMU-system

The SAMU-system was developed in the late 1990s with functionality and features adapted to then existing requirements and organisation at SCB. During the years though, users has gained experience and knowledge meaning that limitations with the system have been identified. A major limitation is that the features in the SAMU-system are built in and cannot be used as "stand alone" tools. The user is obligated to go through step 2) - 7) in the list above in order to draw a coordinated sample from one SF-version. Even though the frame population and sampling design in most cases already has been realised in the test phase. Thus, the user does not consider the SAMU-system as a technical support, rather as a limitation because the built in features are not user-friendly. The non-optional use of a SF, choice of method for sample coordination and number of available sampling designs has also been seen as limitations. Furthermore, since the SAMU-system was developed, a process-oriented organisation has been implemented at SCB and in connection with this the responsible for the SAMU-system was spread out on three different departments. Consequently, the responsibility for the different parts in the SAMU-system has not been quite clear.

3 A generic statistical production environment

During the last years, SCB has focused on standardisation of the statistical production process striving for a standardised set of methods and software's constituting a general toolbox. In other words, SCB is moving to a generic statistical production environment and

one of the basic ideas behind such an environment is that one and the same tool can be used for a specific process, for example sample selection. Moreover, SCB has decided to move from VB6 to .net, which means that the current SAMU-system is outdated. Altogether, this meant an opportunity to develop a completely new tool for sample selection without the limitations experienced from the SAMU-system. As a result, a new generic tool for sample selection is under development, possible to use for all sorts of frame populations and samples: coordinated, independent, multistage and multiphase drawn from any kind of register. This tool should also be possible to use for drawing samples during the test phase and for studies based on large numbers of replicates of samples (simulation studies). Furthermore, this tool should be a tool *isolated* for sample selection. This enables the development of a flexible, user-friendly and effective (high-performance) tool for sample selection. In addition, responsibility for administration, maintenance and development of this tool can be clearly identified (in contrast to responsibility in the SAMU-system).

4 A new tool for sample selection

4.1 General information about the new sample selection tool

The sample selection tool is developed in the software SAS and consist of several SAS-macros, more or less one macro per sampling design. The very basic idea is that this tool should be completely user supplied meaning that it will be possible to use it in any SAS-environment. In other words, a dataset including the frame population for the specific survey together with the sampling design and other necessary information for the sample selection must be provided.

Surveys included in the economical-statistical system must of course provide a frame population based on one version of the SF (with the associated PRNs). Note that the provided PRNs must be transformed in such a way that the first *n* units to the right of zero always are selected (independently of which block the survey actually is placed in). Surveys outside the economic-statistical system can be coordinated by providing PRNs (unique for a specific survey or from one the SFs). The tool generates random numbers when PRNs are missing.

4.2 *Sampling designs covered by the new tool*

In this first version of the sampling selection tool are the following sampling designs covered (all of them use a sequential sampling scheme and can be used with stratification):

- Simple random sampling
- Systematic sampling
- Pareto πps sampling
- Sequential poisson sampling
- Poisson mixture sampling
- Pareto mixture sampling
- Cluster sampling in two stages

These sampling designs were selected because they are used at SCB (more or less common) and because they have fix sample size and are possible to combine with coordination by PRNs (except systematic sampling).

Name of the dataset including the frame population and selected method must always be provided as input to the tool, other parameters needed depends on chosen method. Output from the tool is a SAS-dataset: a copy of the frame population complemented with an

indicator on whether the unit is included in the sample or not, inclusion probabilities and sampling weights. The following parameters (variable names in the dataset containing the frame population) must/can be given:

(Stratified) simple random sampling (ST)SRS
Stratum, *sample size* and *PRN* (optional)

(Stratified) Pareto πps sampling (ST)PAR
Stratum, *sample size*, *size measure*, *PRN* (optional) and *lambda* (inclusion probabilities, optional). Note that the tool handles the case with units where $\lambda_k > 1$ and makes the necessary recalculations until all units included in the sample selection has $\lambda_k < 1$. By providing inclusion probabilities (λ_k) can the user, for example, define the completely enumerated part as all units with $\lambda_k > 0.99$.

(Stratified) systematic sampling (ST)SY
Stratum, *sampling interval* (or *sample size*) and *starting point* (optional). Note that by providing sample size the tool calculates the sampling interval as the sampling fraction.

(Stratified) Poisson mixture sampling, (ST)POMIX
Stratum, *sample size*, *size measure*, *PRN* (optional), *lambda* (inclusion probabilities, optional) and *B* (Bernoulli width). Note that Sequential poisson sampling is obtained for B=0 (see Kröger et al., 2003).

(Stratified) Pareto mixture sampling, (ST)PARMIX
Stratum, *sample size*, *size measure*, *PRN* (optional), *lambda* (inclusion probabilities, optional) and *B* (Bernoulli width). Note that Pareto πps sampling is obtained for B=0 (see Kröger et al., 2003).

Cluster sampling in two stages
Cluster variable, *method1*, *method2*, *number of cluster* or *sampling interval*. SRS and SY can be used in the first stage (method1). When SRS is used in the first stage must number of cluster be provided, otherwise sampling interval. In the second stage (method2) can either (ST)SRS, (ST)SY, (ST)PAR, (ST)POMIX or (ST)PARMIX sampling design be used. For method 2 must the same parameters be provided as in ordinary element sampling.

5 Future and discussion

5.1 Future

A first version of the sampling selection tool is finalised and released for testing at SCB. Thereafter an implementation in the general toolbox is planned. Already known is a need for development to cover two-phase sampling and systematic πps sampling. Implementation of systematic πps sampling will also offer a better solution to the problem with non-integer sampling interval in simple (ST)SY compared to the one currently implemented. Poisson and Bernoulli sampling would also be of interest to implement in the near future. Moreover, if proven worthwhile, this tool could perhaps be developed to support sample selection in ordinary panel designs (coordination by sample panels) and/or other coordination technique(s).

5.2 Discussion

This paper presents a new tool for sample selection as well as points out certain areas related to sample selection but not covered by the new sample selection tool. It would be most valuable for SCB to have some knowledge about the situation concerning sample selection (and areas related to sample selection) at other NSI:s participating in EESW17. For example, concerning the following items:

- existing general tool/software for sample selection
- tools for the other steps related to sample selection (like frame creation, allocation of sample sizes etc.)
- database for general and unified storage of frame populations and samples
- administration of surveys in general
- administration of surveys within the economical-statistical system
- organisation, responsibility for administration, development and maintenance for sample selection tools/software

6 References

Kröger, H., Särndal, C.E. and Teikari, I. (2003). Poisson Mixture Sampling Combined with Order Sampling. Journal of Official Statistics, 19, 59–70.

Lindblom, A. (2003). SAMU - The system for coordination of frame populations and samples from the Business Register at Statistics Sweden, Background Facts on Economic Statistics 2003:3, Statistics Sweden.

Ohlsson, E. (1992). SAMU – The System for Co-Ordination of Samples from the Business Register at Statistics Sweden: A Methodological Description. R&D Report 1992:18. Statistiska centralbyrån.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers. I Business Survey Methods (red. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge och P. Kott). New York: Wiley, 153–169.

Rosén, B. (2000). *A User's Guide to Pareto πps Sampling.* R&D Report 2000:6. Statistiska centralbyrån.