The Simulation Study of Survey Cost and Precision

Mārtiņš Liberts[1]

[1]University of Latvia, e-mail: martins.liberts@gmail.com

Abstract

Cost efficiency is a desirable property for sample surveys done in practice. It is a common task for a statistician to find the balance between precision and cost during the planning stage of a survey. For example, cluster sampling can be preferable choice regarding cost efficiency because the reduction of cost can dominate the loss in precision. Assessment of survey design regarding cost efficiency can be complex task. The approach presented in the talk is to use the methodology of simulation experiments as a tool for the cost efficiency analysis. Artificial population data is used in simulation experiments. The artificial population is created using the data from the Population Register of Latvia and the data from Latvian Labour Force Survey (LFS). The analysis of sampling design used for Latvian LFS will be presented. Two stage sampling design is used for the Latvian LFS where census counting areas are primary sampling units and dwellings are secondary sampling units. The design will be compared with other traditional sampling designs regarding cost efficiency.
Keywords: Survey sampling, simulation, cost, precision

1 Introduction

The idea of the study comes from purely practical necessity. National statistical institutes (NSI) usually are the main providers of the official statistics. The customers of the official statistics are society or tax payers in other words. Cost efficiency is one of the very desirable property for the government spendings. Somebody can ask a question – is the survey organised by NSI is cost efficient?

The aim the of the study is to develop a practical tool to compare different sampling strategies regarding a cost efficiency. The work is based on the Labour Force Survey.

2 The target population

The target population of the Labour Force Survey (LFS) usually is defined as all residents permanently living in private households (age group of the working age – 15-74 is the main domain of the interest). The target population is constantly changing over time. The target population is observed on weekly bases by the methodology of LFS (European Communities, 2003). Questioning of all residents every week would be required if LFS would be done as a census (full survey of whole target population). The example of the LFS target population is given by the table 1.

The population can be represented as a table with rows representing individuals and columns representing weeks. There are N individuals labelled with labels $1, 2, \ldots, N$. The populations in the table 1 refers to W weeks. Weeks are labelled with labels $1, 2, \ldots, W$.

There is an assumption of fixed set of individuals during the period of W weeks. It means there is not any "birth" or "death" during the time period under consideration. This assumption does not hold in practice.

Table 1: The target population of LFS

i	w=1	w=2	w=3	w=4	w=5	$\cdots$	w=W
1	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$\cdots$	$y_{1,W}$
2	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$\cdots$	$y_{2,W}$
3	$y_{3,1}$	$y_{3,2}$	$y_{3,3}$	$y_{3,4}$	$y_{3,5}$	$\cdots$	$y_{3,W}$
4	$y_{4,1}$	$y_{4,2}$	$y_{4,3}$	$y_{4,4}$	$y_{4,5}$	$\cdots$	$y_{4,W}$
5	$y_{5,1}$	$y_{5,2}$	$y_{5,3}$	$y_{5,4}$	$y_{5,5}$	$\cdots$	$y_{5,W}$
6	$y_{6,1}$	$y_{6,2}$	$y_{6,3}$	$y_{6,4}$	$y_{6,5}$	$\cdots$	$y_{6,W}$
$\cdots$							
N	$y_{N,1}$	$y_{N,2}$	$y_{N,3}$	$y_{N,4}$	$y_{N,5}$	$\cdots$	$y_{N,W}$

3 Parameters of interest

Two population parameters are considered – total and ratio of two totals.

3.1 Total

Weekly total for the week w is defined by the equation 1.

$$Y_w = \sum_{i=1}^{N} y_{i,w} \tag{1}$$

Quarterly total for the quarter q is defined by the equation 2. There is an assumption – all quarters consist of 13 weeks. There are some quarters with 14 weeks in real calendar.

$$Y_q = \frac{1}{13} \sum_{w=j}^{j+12} Y_w = \frac{1}{13} \sum_{w=j}^{j+12} \sum_{i=1}^{N} y_{i,w} \tag{2}$$

Yearly total for the year y is defined by the equation 3. There is an assumption – all years consist of 4 quarters or 52 weeks. There are some years with 53 weeks in real calendar.

$$Y_y = \frac{1}{4} \sum_{q=k}^{k+3} Y_q = \frac{1}{52} \sum_{w=j}^{j+51} Y_w = \frac{1}{52} \sum_{w=j}^{j+51} \sum_{i=1}^{N} y_{i,w} \tag{3}$$

3.2 Ratio of two totals

Weekly ratio of two totals for the week w is defined by the equation 4.

$$R_w = \frac{Y_w}{Z_w} = \frac{\sum_{i=1}^{N} y_{i,w}}{\sum_{i=1}^{N} z_{i,w}} \tag{4}$$

Quarterly ratio of two totals for the quarter q is defined by the equation 5.

$$R_q = \frac{Y_q}{Z_q} = \frac{\sum_{w=j}^{j+12} Y_w}{\sum_{w=j}^{j+12} Z_w} \tag{5}$$

Yearly ratio of two totals for the year y is defined by the equation 6.

$$R_y = \frac{Y_y}{Z_y} = \frac{\sum_{q=k}^{k+3} Y_q}{\sum_{q=k}^{k+3} Z_q} = \frac{\sum_{w=j}^{j+51} Y_w}{\sum_{w=j}^{j+51} Z_w} \tag{6}$$

4 Design efficiency

4.1 The balance of variance and cost

Assume an arbitrary population parameter θ. Assume there is a probability sample s drawn by known sampling design $p(s)$. θ can be estimated using an estimator $\hat{\theta}_p$. The variance of $\hat{\theta}_p$ is denoted by $V\left(\hat{\theta}_p\right)$.

Assume a cost associated to a sample s. This is a cost what survey organiser has to spend to carry out the survey with sample s. A cost can be expressed in money, time or other quantity. Assume there is a cost function $c(s)$. The cost of sample s can be computed by the cost function $c_s = c(s)$. c_s is a random because s is a random sample. The expectation of c_s under sampling design $p(s)$ is notated as $E(c_s) = C_p$.

Usual desire is to minimise $V\left(\hat{\theta}_p\right)$ and C_p. Unfortunately these are conflicting tasks. You have to increase cost to reduce the variance and variance goes up when cost is reduced. The usual task of statistician is to construct sampling design $p(s)$ so that C_s and $V\left(\hat{\theta}_p\right)$ would be in "balance".

4.2 Design effect

There is a need for a measure of design efficiency. Assume two sampling designs:

- Simple random sampling – srs
- Alternative sampling design – $p(s)$

The classical design effect is a ratio of variances under condition of equal sample sizes defined by the equation 7.

$$deff\left(p, \hat{\theta}, n\right) = \frac{V\left(\hat{\theta}_p | E(n_p) = n\right)}{V\left(\hat{\theta}_{srs} | n_{srs} = n\right)} \tag{7}$$

$\hat{\theta}_p$ denotes π estimator under sampling design p(s), $\hat{\theta}_{srs}$ denotes π estimator under simple random sampling.

Alternative design effect can be introduced by the equation 8. It is defined as a ratio of variances under condition of equal expected costs.

$$deff^{\star}\left(p, \hat{\theta}, \gamma\right) = \frac{V\left(\hat{\theta}_p | C_p = \gamma\right)}{V\left(\hat{\theta}_{srs} | C_{srs} = \gamma\right)} \tag{8}$$

The design effect defined by (8) could be used as a measure of design efficiency. Assume two sampling designs – $p(s)$ and $q(s)$.

Definition 1 *The sampling design $p(s)$ is more efficient then the sampling design $q(s)$ for estimation of θ with survey budget γ if $deff^{\star}\left(p, \hat{\theta}, \gamma\right) < deff^{\star}\left(q, \hat{\theta}, \gamma\right)$.*

The definition 1 is equivalent to the definition 2:

Definition 2 *The sampling design $p(s)$ is more efficient then the sampling design $q(s)$ for estimation of θ with survey budget γ if $V\left(\hat{\theta}_A, C_A = \gamma\right) < V\left(\hat{\theta}_B, C_B = \gamma\right)$.*

5 Simulation

Design efficiency can be measured with help of simulation experiments. Artificial population data are necessary to carry out the simulation experiments. The artificial population data are created from the data of the Latvian Population Register and the survey data of Latvian LFS.

5.1 Sampling designs

The task of this research is to measure the design efficiency for three sampling designs used to select a quarterly sample – sample for 13 weeks.

There are two questionnaires for LFS – household questionnaire and individual questionnaire.

5.1.1 SRS of individuals

The SRS of individuals is selected. Sample of individuals is allocated randomly and evenly over 13 weeks.

The sample of dwellings is constructed from the sample of individuals. A dwelling is sampled if at least one dwelling member is sampled. A household questionnaire is filled for each sampled dwelling. An individuals questionnaire is filled for each sampled individual.

5.1.2 SRS of dwellings

The SRS of dwellings is selected. Sample of dwellings is allocated randomly and evenly over 13 weeks.

A household questionnaire is filled for each sampled dwelling. Individuals questionnaires are filled for all individuals from a sampled dwelling.

5.1.3 Two stage sampling design

This is the sampling design used in practice for Latvian LFS (Liberts, 2010).

A household questionnaire is filled for each sampled dwelling. Individuals questionnaires are filled for all individuals from a sampled dwelling.

5.2 Cost function

The cost is expressed as time necessary for field interviewers to carry out the survey in the simulation. There two components:

- Time for travelling $t_1(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}}$ where G is a number of interviewers, d_g is a distance done by interviewer g to carry out the survey, $\bar{v}$ – an average travelling speed of interviewer.
- Time for interviewing $t_2(s) = m \cdot \bar{t}_H + n \cdot \bar{t}_P$ where m is number of dwellings taking part in survey, n is the number of individuals taking part in survey, $\bar{t}_H$ is an average time for a household interview, $\bar{t}_P$ is an average time for a personal interview.

The following cost function is used to measure the cost of the survey:

$$c(s) = t_1(s) + t_2(s) = \frac{\sum_{g=1}^{G} d_g}{\bar{v}} + m \cdot \bar{t}_H + n \cdot \bar{t}_P \tag{9}$$

6 Results

The results of the simulation study will be presented during the workshop.

References

European Communities (2003). *The european union labour force survey.* Office for Official Publications of the European Communities, Luxembourg.

Liberts, M. (2010). *Official statistics – methodology and applications in honour of Daniel Thorburn*, chap. The Redesign of Latvian Labour Force Survey. The Department of Statistics, Stockholm University (in collaboration with Statistics Sweden), Stockholm, Sweden, pp. 193–203.