Combinining samples of HBS and EU-SILC in Estonia

Julia Aru[1]

[1]University of Tartu, Statistics Estonia, e-mail: julia.aru@gmail.com

Abstract

Statistics Estonia conducts several social surveys with very similar or identical target populations. These surveys focus on different topics, but still contain a block of common questions. Thanks to that, it is possible to combine samples of different surveys to produce more detailed output on these common questions and increase precision. We discuss combining of the samples of two surveys: Survey on Income and Living Conditions (EU-SILC) and Household Budget Survey (HBS). First is a longitudinal survey with rotational design, and second is a purely cross-sectional survey, so the main challenge is computing weights for the combined sample. We discuss several approaches to weighting, gain in precision for combined sample, and final recommendation of the weighting method.
Keywords: Workshop, template, contributed paper

1 Introduction

Statistics Estonia, like any other national statistical office, conducts a lot of social surveys. These surveys focus on various topics and may differ in terms of target population and design, but there is always some overlap in questionnaires, e.g. education, socio-economic status, living conditions etc. In this situation, estimates for common questions can be derived from several surveys. By combining samples of several surveys and estimating common questions from the bigger combined sample, NSI can avoid publishing different estimates for the same phenomenon, which is confusing for users, as well as increase precision of output. This approach is already used for years in several European countries, like Netherlands and UK, while this article will describe the results of the first exercise of this kind in Statistics Estonia.

Two surveys are used in this analysis: Household Budget Survey (HBS) and Survey on Income and Living Conditions (EU-SILC). Features of these surveys are summarized in Table 1.

Table 1: HBS and EU-SILC

Feature	HBS	EU-SILC
Target population	All persons in private households	
Longitudinal component	No longitudinal component, purely cross-sectional	Household remains in the sample for 4 years; a sample of any single year consists of a new part, which is approached for the first time, and repeated part, which is

		approached for the second, third or fourth time.
Sampling design	Stratified systematic sampling of persons from the Population Register, with whole household included along with selected person, which results in PPS sampling for households;	
Non-response correction	Logistic regression with age and gender of selected person, county group and degree of urbanisation.	Logistic regression for new and repeated part separately. Predictors in new part: age and gender of selected person, county group and degree of urbanisation. Predictors in repeated part: tenure status, type of household, county group, nationality, degree of urbanisation, income decile in previous year.
Calibration	Gender-age group and county	
Sample size in 2010	3600 hhs	5000 hhs

The main challenge in this context is computation of weights for combined sample. There are survey-specific weights, which account for the design of the survey, are corrected for non-response and calibrated. Simple method of weighting uses these existing weights after adjusting with a scaling factor as will be described below. More complicated method in some sense starts from the beginning and calculates the probabilities to be included into the combined sample for each household. In the following section we will describe each method in more detail.

2 Methods for weighting the combined sample

2.1 Method of adjusting existing weights

The following method of calculating weights uses existing survey-specific weights and thus is quite simple to implement as the only thing an analyst needs to calculate is an adjustment factor. This method is also referred to as the method on combining samples by Iachan *et al.* (2003) and O'Muircheartaigh & Pedlow (2002). It is simpler to explain to users, more transparent and less dependent on models and assumptions. Nevertheless, it is not clear how applicable it is in case of differences of target populations between the surveys, as will be discussed later.

In general, when adjusting existing weights, we need to calculate a factor $\alpha \in [0,1]$ by which we multiply the weights of the first survey. The weights for the second survey are then multiplied by the factor $1-\alpha$ to ensure that weights for the whole combined sample still sum up the population size. A variety of methods has been proposed for calculating α (see, for example Korn & Graubard, 1999). We use the method described in O'Muircheartaigh & Pedlow (2002), which exploits samples sizes and variability of the survey-specific weights:

$$\alpha = \frac{n_1 / d_1}{n_1 / d_1 + n_2 / d_2}, \text{ where } d_1 = 1 + \frac{Var(w_i, i \in SILC)}{\overline{w}_{SILC}^{\,2}},\ d_2 = 1 + \frac{Var(w_i, i \in HBS)}{\overline{w}_{HBS}^{\,2}},$$

and n_1 and n_2 are respectively EU-SILC and HBS sample sizes.

For the two surveys used in this analysis, $\alpha = 0.561$. Quantities d_1 and d_2 are well-known expressions for the

design effect (the part of that due to the variability of weights), and thus n_1/d_1 and n_2/d_2 are effective sample sizes of the surveys. The survey with larger effective sample size receives bigger factor and thus is more influential on the estimates.

Finally, weights were calibrated by 5-year gender-age groups and county.

2.2 Method of cumulating probabilities

Another approach to weighting is to calculate the probability to be included (and respond) in the combined sample. Here, for each household, we calculate the probability to be included in the combined sample as a whole, i.e. to be included in one of the samples, independently on which it was really included in. So, for example, for a household from HBS we need to calculate the probability that it would have been included in EU-SILC, and vice versa, taking into account survey-specific response pattern.

Because of different response models used for different parts of EU-SILC, this survey is divided into the new part and repeated part. In what follows we treat the combined sample as the concatenation of three (instead of two) surveys: EU-SILC repeated part, EU-SILC new part, HBS.

For this method to be comparable with simple method we use the same non-response adjustment methods as described in Table 1. For EU-SILC repeated part, probability to respond in 2010 is modelled as the product of probability to respond in the first year (in the year of first selection into the sample) and probability to respond in 2010 (given the household has responded in the year of selection).

We use following notation:

S – combined sample;

R – response set for the combined sample;

S_1, R_1, R'_1 – respectively the sample, first year response set and 2010 response set for EU-SILC repeated part

S_2, R_2 – sample and 2010 response set for EU-SILC new part;

S_3, R_3 – sample and 2010 response set, HBS;

As $R = R'_1 \cup R_2 \cup R_3$ and surveys are negatively coordinated (households already participating in one of the surveys are dropped prior to the sample selection for the other) the probability of household i to be included into the combined sample and respond can be written as:

$$\begin{aligned} \Pr(i \in R) &= \Pr(i \in R'_1) + \Pr(i \in R_2) + \Pr(i \in R_3) = \\ &= \Pr(i \in R'_1 \mid i \in R_1) \Pr(i \in R_1 \mid i \in S_1) \Pr(i \in S_1) + \\ &+ \Pr(i \in R_2 \mid i \in S_2) \Pr(i \in S_2) + \Pr(i \in R_3 \mid i \in S_3) \Pr(i \in S_3). \end{aligned} \quad (1)$$

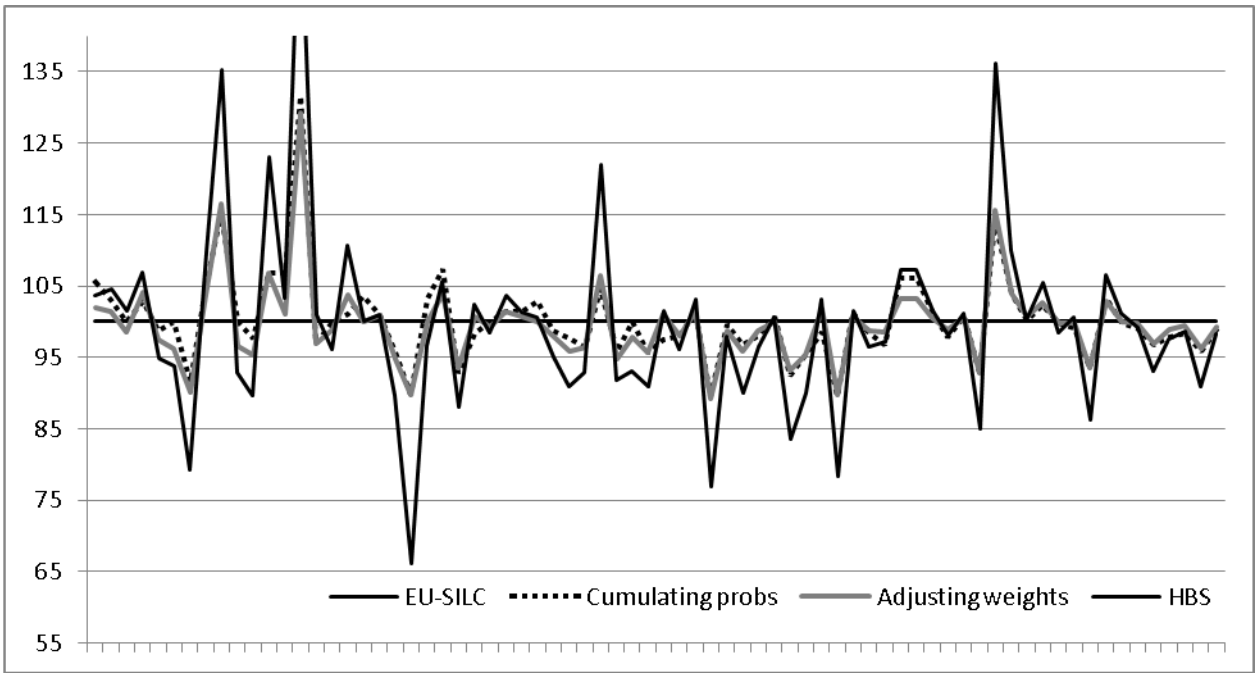
That is, for example, for every household in HBS we need to calculated the inclusion probability and response probability *as if* it is included in EU-SILC new part and *as if* it is included in EU-SILC repeated part. To calculate response probabilities, three response models had to be fitted to the survey-specific data (logistic regression) as shown in Table 1. Fitted logistic regression equation was then applied to every household to calculate response probability, irrespective of the survey the household originated from.

Preliminary weight is a reciprocal of probability (1), and before use it was calibrated by 5-year gender-age groups and county.

3 Comparison of estimates

In spite of different weighting methods for the combined sample, estimates of variables measured in both surveys were very similar. Figure 1 shows a number of estimates calculated with each of the two methods described above as well as the same estimates from specific surveys. All estimates are calculated relative to the EU-SILC estimate, taken as 100%.

Figure 1: Estimates of common variables calculated with different weighting methods



We calculated variance estimates and coeficients of variation for common variables with each of methods, and compared estimates of some parameters with the true values from registers to assess bias. Results are shown in Table 2.

Table 2: Comparison of estimates

	EU-SILC	Cumulating probabilities	Adjusting weights	HBS
Design effect	1,69	1,54	1,66	1,57
Sample size	4972	8604	8604	3632
Effective sample size	2947	5569	5180	2306
Average cv of estimates (%)	3,10	2,28	2,35	3,56
Average absolute relative bias of estimates (%)	17,2	13,5	13,5	21,4

Method of cumulating probabilities seems to give less variable weights, which gives some gain in precision of estimates as compared to the method of adjusting weights. But both methods decreased bias equally and gain in precision appears to be marginal.

4 Summary and future plans

The first exercise on combining sample of two surveys gave very promising results. We compared two methods of weighting: adjusting existing weights and cumulation probabilties. Methods gave very similar results both in terms of precision and bias. So, at least for the two surveys examined, we can use a simpler method of adjusting existing weights. Method of cumulating probabilities is much more difficult to implement, it requires calculating design inclusion probabilities from the beginning and fitting of several response models on different sets of data. For the two surveys at hand, it gave minor gain in precison as compared to the other method, so we suppose it is not worth the effort in future. Statistics Estonia is planning to use the combined sample for producing regular statistics from 2013, with method of adjusting weights as we recommended.

Still, we are going to continue reseach on this topic. It is not clear, would simpler method perform so well in the case of more serious differences between the designs and target populations of the surveys involved. Next step would be to add the Labour Fource Survey (LFS) to the two surveys used in this analysis. LFS has somewhat different design and target population and we are going to repeat this comparison of weighting methods and give recommendations on the weighting methods for combination of three surveys: EU-SILC, HBS and LFS.

Another challenging topic is to re-calibrate survey-spesific weights to the estimates of combined sample. This is also appealing since we don't know much about first year non-responders in our surveys (available information is limited to register variables such as place of residence, age and gender). For example, with combined sample we could estimate the distribution by education status more precisely (education is considered to be a good predictor for many other topic variables) and then re-calibrate each survey by education. This is expected to improve precision of survey-specific estimates, but till then remains a topic for future research.

References

Iachan, R., Saaverda, P. & Robb, W. (2003) Two weighting schemes for combining sample in the Health Behaviors in School-age Children Survey. *2003 Joint Statistical Meetings - Section on Survey Research Methods*

Korn, E. L. & B. I. Graubard (1999). *Analysis of Health Surveys*. Wiley, New York.

O'Muircheartaigh, C. & Pedlow, S. (2002). Combining samples vs. cumulating cases: a comparison of two weighting strategies in NLSY97. *ASA Proceedings of the Joint Statistical Meetings*, pp. 2557-2562.