ADAPTIVE SAMPLE SURVEY DESIGN IN DATA COLLECTION

J.Voronova[1]

[1] Central statistical bureau of Latvia, Latvia
e-mail: jelena.voronova@csp.gov.lv

Abstract

The responsive adaptive survey design (ASD), utilizing R-indicators as measures of representativeness, is tested in Central statistical bureau of Latvia (CSB) as a flexible approach for organizing social surveys. R-indicators help to identify potential bias by measuring the degree of difference between responding and non-responding sample groups. Based on the monitoring and analyses of the R-indicators, active interventions are implemented during data collection process to increase the chances of obtaining a representative set of final response unit, thereby reducing variance in the weights of the final survey data.

Using the notation and definition of response propensities as set out in Schouten, Cobben and Bethlehem (2009) and Shlomo, Skinner and Schouten (2012), denote U the set of units in the population U=1,2,..,i,..N and s the set of units in the sample s=1,2,..,i,..n . Denote a response indicator variable R_i which takes the value 1 if unit i in the population responds and the value 0 otherwise. The response propensity is defined as the conditional expectation of R_i given the vector of values x_i of the vector X of auxiliary variables:

$$\rho_x(x_i) = E(R_i=1|X=x_i) = P(R_i=1|X=x_i) \quad (1)$$

and also denote this response propensity by ρ_x .

Define the R-indicator as:

$$R(\rho_x) = 1 - 2S(\rho_x) \quad (2)$$

Estimation of the response propensity is based on logistic regression model and estimator of the variance of the response propensities:

$$\hat{S}^2(\hat{\rho}_x) = \frac{1}{N-1}\sum_s d_i(\hat{\rho}_x(x_i) - \overline{\hat{\rho}_x})^2 \quad (3)$$

where $d_i = \pi_i^{-1}$ is the design weight or inverse inclusion probabilities and $\overline{\hat{\rho}_x} = \frac{1}{N}\sum_s d_i \hat{\rho}_x(x_i)$. Thereby, estimation of the R-indicator $\hat{R}(\hat{\rho}_x) = 1 - 2\hat{S}(\hat{\rho}_x)$.

As in variance analysis, R-indicator has the same characteristics and could be split into unconditional partial indicators, which measures the distance to representative response for single auxiliary variables and are based on the between variance given a stratification with categories of Z and conditional partial R-indicators measure the remaining variance due to variable Z within sub-groups formed by all other remaining variables as in Schouten, Shlomo and Skinner (2011).

Survey responsive data collection design concept was piloted in CSB on three person surveys all of them was conducted by using systematic stratified simple random sample:

- Objective of the survey “Mobility of Latvian population in 2021” (MOBS) is to find out the mobility habits of the population. 8 978 persons aged 15 to 84 years living in private households in Latvia selected into sample, the response rate in the survey accounted for 60.4 %.

The survey took place at time when the spread of COVID-19 had particularly intensified and stricter restrictions were introduced in Latvia in order to reduce this spread.

- First "Survey on Gender-Based Violence" (SGBV) 2021 is aimed at collecting information on prevalence of various types of violence in Latvia based on common methodology developed by the Eurostat. SGBV covers personal safety and experience with unwanted behaviour at work, in society, partnership, family, and childhood. The target population of the survey covers people aged 18–74 living in private households in Latvia. Within the framework of the survey, 6 300 people were interviewed. The survey was conducted during rapid spread of COVID-19 and strict restrictions imposed to fight it.
- Adult Education Survey (AES) 2022 is aimed at acquiring internationally comparable data on adult participation in lifelong learning activities – formal education, non-formal education and training, informal learning. The questions covered participation in education activities within the last 12 months. Target population of the survey starting from 2022 cover people aged 18–69 living in private households - 8764 usually residents of Latvia. Answers to the questionnaire questions developed by the CSB were given by 5 492 persons.

The focus of the ASD approach was set on ensuring the quality of fieldwork, with a particular emphasis on the representativeness of sample's response unit set. Several steps were taken during fieldwork to achieve the goal. At the first part of the data collection, R-indicators were used for monitor needs, afterward the groups of imbalance were identified and resources of interviewers were redirected to data collection of those groups.

Response propensity model was developed for each monitoring date during data collection period. The response propensities were estimated a generalized linear model (GLM), a generalization of the classical linear model, with the binomial family logistic-regression model (logistic link function). The set of auxiliary variables were built from social-demographic variables and paradata. Various approaches were used for variable selection, including correlation analysis, evaluation of the amount of available data, level of explanation of the propensity to respond. Individual final set was evaluated for each survey. Selection of the final model specification was evaluated by the automatic *stepAIC* procedure from the *MASS* package (Venables, W. N. & Ripley, B. D. 2002), thus iteratively were reviewed all possible models from the initially passed parameters and left only those variables where the AIC criteria was the smallest.

There were no pre-defined methods for ASD, CSB usually uses multi-mode research method, but the impact of COVID-19 was still significant in 2021. The cancellation of face-to-face interviews led to shift to CATI in 2020. It the situation of a defined fieldwork period, a limited resources as number of interviewers were available, at least one contact for every sample unit were allowed. All resources were planned to be redirected to imbalanced groups.

An important implication of the study was individuality of the survey aim and scope, its influence on the results of the tests. Although more active intervention was made in MOBS, the representativeness of the response set increased in both (MOBS and SGBV) at the end of data collection period. In association with assessed results, possible assumption is the survey aim and type of questions affects representativity - MOBS questionnaire is about habits in specific period, while SGBV survey questions is more about whole life experience.

One of the aims of data processing was to assess variance, and the analysis showed better results in MOBS than SGBV, mostly because of different goals of the surveys. MOBS data, before and after re-directing interviewer resources, showed a reduction of variance. Additionally, an overestimation of the variable of interest was observed in the imbalanced response set. SGBV showed an imbalance in the final response set by sex, and the impact of COVID-19 restrictions on survey results were also observed.

Some valuable lessons were learned in organization and managing ASD in CSB during the 2021 surveys. An ASD dashboard for monitoring needs, which was evaluated by the survey manager, was introduced in AES 2022. Process of results analysis of ASD is ongoing, and the results will be available later this summer.

Keywords: Nonresponse, representativeness, R-indicator, adaptive design

References

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schouten B., Cobben F., Bethlehem J. (2009) *Indicators for the representativeness of survey response,* Computer Science, Chemistry Dalton Transactions.

Schouten B., Peytchev A., Wagner J. (2018) *Adaptive survey design*, Chapman & Hall/CRC Statistics in the Social; Behavioral Sciences.

Schouten B., Shlomo N. (2015*) Selecting adaptive survey design strata with partial R-indicators,* CBS, https://www.cbs.nl/-/media/imported/documents/2015/51/2015-selecting-adaptive-survey-design-strata-with-partial-r-indicators.pdf?la=nl-nl

Schouten B., Shlomo N., Skinner C.J. (2011) *Indicators for monitoring and improving representativeness of response*, Journal of Official Statistics, Vol. 27, No. 2, 231-253.

Shlomo N., Schouten B., de Heij V. (2013) *Designing adaptive survey designs with R-indicators*, NTTS 2013, https://ec.europa.eu/eurostat/cros/system/files/NTTS2013fullPaper_63.pdf

Shlomo N., Skinner C., Schouten B. (2012) Estimation of an indicator of the representativeness of survey response, Volume 142, Issue 1, January 2012, 201-211.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0