Representativity indicators for business surveys based on population totals

Pim Ouwehand, Barry Schouten, Vincent de Heij
Department of Methodology and Quality, Statistics Netherlands

August 21, 2009

Abstract

The RISQ (Representativity indicators for survey quality) project aims to develop quality indicators for survey response. Whereas the response rate alone only measures the size of the response, representativity indicators also measures the quality of the response and therefore the potential impact of the non-response. These representativity indicators (*R-indicators*) measure to what extent the response resembles the population as a whole, and thus how good estimates of population characteristics will probably be.

1 Introduction

Business surveys are carried out in order to obtain estimates of unknown population characteristics. Also administrative registers can be used to obtain these estimates. Registers are different from surveys in that they aim at a complete enumeration of a population. In many surveys and registers part of the data is missing due to non-response and time-delays in reporting of businesses to register holders. This can have a significant impact on the quality of statistics based on these sources of data. The available data may not resemble the population and lead to a bias in estimates. If we want to describe the survey quality, the response rate alone is not a good indicator, and thus other indicators are needed. The RISQ (Representativity Indicators for Survey Quality) project[1] aims to develop and test indicators that measure the degree to which the response of a survey or available data in a register resembles the population. These indicators may serve three goals: to enable the comparison of different surveys or of a single survey in time, to assist in monitoring survey data collection, and most ambitiously to guide survey data collection. Indicators may also be of help in monitoring and evaluating the completion of a register. The application to registers is especially useful when registers have a time lag and are filled gradually in time.

A low response rate is often used as an indicator of low survey quality. Although a low response rate means a reduced sample size and may therefore lead to a larger estimation error, it will not necessarily cause estimates to be biased. If, for example, all elements have the same response probability, estimates remain unbiased. We need a measure that describes the quality of the response, by comparing the characteristics of the response to those of the non-response.

However, whether the response is different from the non-response with respect to the target variable and is thus representative, can hardly ever be established. Only responding elements can be observed. It is the objective of the survey or register to observe

[1]The RISQ project, funded by the European 7th Framework Programme, is a joint effort of the NSI's of Norway, The Netherlands and Slovenia, and the universities of Leuven and Southampton to develop quality indicators for survey response

characteristics of households and enterprises that are not known. If auxiliary information is available, one can indirectly measure whether response and non-response differ. It is the aim of the RISQ project to use auxiliary information to develop representativity indicators (so called *R-indicators*).

The auxiliary data will form the reference with respect to which we will measure the representativity. We can compare representativity directly to the sample, that is, link sample units directly to auxiliary information from other sources. In this project however, we will compare representativity directly to the population, i.e. the survey must contain questions for which population totals are known from other sources.

The main objectives of this project are to develop representativity indicators, derive their theoretical properties, and perform and empirical validation of them. The indicators can be used to compare the representativity of the response in time and between surveys and registers.

2 Definition of representativity

There are several definitions of representativity. The concept can be linked to statistical estimators, so that a response can be considered representative for a statistic it allows for estimation of that statistic without non-response error. Here, however, we restrict ourselves to the characteristics of the response alone, and adopt the view that representativity corresponds to the difference in characteristics between respondents and non-respondents.

Two aspects are important here. First, our definition is dependent upon information on auxiliary variables. Hence, any indicator will have to be disseminated together with a statement about what auxiliary information was employed to evaluate representativity. Second, our representativity indicator is estimated from sample data and so suffers from loss of precision as the sample size decreases.

In order to define a representativity indicator, one approach is to use response probabilities as the basis for representativity indicators. However, a representativity indicator that uses response probabilities cannot be computed in practice. The values of the response probabilities are unknown. Therefore they have to be estimated using some kind of model (logit, probit or linear). But, estimating the response probability for every element in the population requires knowledge of the values of the predictor variables for every element. Such information is often not available, so that one has to rely on sampling information only.

Since using response probabilities (denoted by ρ_k) is not straightforward, we restrict ourselves to response propensities (denoted by $\rho_k(X)$). The response propensity is the response probability given the values of some auxiliary variable(s). They are also unknown but can be estimated provided the values of the auxiliary variable(s) are available for both the respondents and the non-respondents. The response propensity of element k (for $k = 1, 2, \ldots, N$ with N the population size) is given by

$$\rho_k(X) = P(R_k = 1|X_k) \tag{1}$$

where $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})'$ is a vector of values of p auxiliary variables. The response R_k assumes the value 1 if element k is selected in the sample and responds, and 0 otherwise. In case of a simple linear model these propensities are estimated from:

$$\rho_k(X) = \sum_{j=1}^{p} X_{kj}\beta_j$$

where β_j are regression coefficients.

One possible R-indicator (Schouten et al., 2008) is based on the variance of the response probabilities. The response probabilities are replaced by response propensities. Then, the

response propensities are estimated using some kind of model. This results in estimates p_k for the response propensities. Next, the R-indicator can be estimated by

$$R(p) = 1 - 2S(p) \tag{2}$$

where

$$S(p) = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(p_k - \bar{p})^2} \tag{3}$$

and

$$\bar{p} = \frac{1}{N}\sum_{k=1}^{N} p_k \tag{4}$$

3 Auxiliary variables and population totals

Earlier work (Schlomo et al., 2009) that was devoted to the construction of R-indicators and the evaluation of their statistical properties assumes that the survey sample is available and can be linked to a set of external data sources like registers and administrative data. The variables from these data sources are known for both respondents and non-respondents to the survey, and function as auxiliary variables for the prediction of response behaviour.

Auxiliary data may be available at the sample level by means of direct linkage to frame data, registrations, and administrative data, but may also be available in aggregated form at the population level only. For example, the only form of auxiliary information that may be available is a set of statistics produced by national statistical institutes. These institutes disseminate tables about a wide range of population statistics, but only at aggregate level. This research is about R-indicators that are based completely on such population statistics and that can be computed without any knowledge about the non-respondents.

We adapt the two indicators developed in (Schlomo et al., 2009) to population tables and population counts. We replace sample covariances and sample means by population covariances and population means. We will call the resulting indicators population-based R-indicators. To our knowledge there is no record in the literature about models for response propensities that employ population information only. In this respect the current paper is innovative and may be applicable to other statistical areas as well. As a consequence, however, we feel this paper is just a first start. More research is likely to be necessary in order to refine estimators and estimation strategies.

References

Bethlehem, J. (2008). Analysing non-response with Anota. *Working paper.* Department of Methodology and Quality, Statistics Netherlands.

Keller, W.J., A. Verbeek, and J. Bethlehem (1984). ANOTA: Analysis of Tables. *CBS-report 5766-84-M1-3.* Department of Statistical Methods, Statistics Netherlands.

Schlomo, N., C. Skinner, B. Schouten, J. Bethlehem, and L.C. Zhang (2009). Statistical properties of representativity indicators. *RISQ deliverable.* RISQ project.

Schouten, B., F. Cobben, and J. Bethlehem (2008). Indicators for the representativeness of survey response. *Survey Methodology.* To be published.

Skinner, C., N. Schlomo, B. Schouten, L.C. Zhang, and J. Bethlehem (2009). Measuring survey quality through representativeness indicators using sample and population based information. *Paper presented at NTTS conference, 18-20 February 2009, Brussels, Belgium.*