Discussion Paper

# Estimation of response propensities and R-indicators using population-level information

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

2016 | 21

Annamaria Bianchi
Natalie Shlomo
Barry Schouten
Damiao da Silva
Chris Skinner

# Content

## Summary

In recent years, there has been a strong interest in indirect measures of non-response bias in surveys or other forms of data collection. This interest originates from gradually decreasing propensities to respond to surveys parallel to pressures on survey budgets. These developments led to a growing focus on the representativeness or balance of the responding sample units with respect to relevant auxiliary variables. One example of a measure is the representativeness indicator, or R-indicator. The R-indicator is based on the design-weighted sample variation of estimated response propensities. It pre-supposes linked auxiliary data. One of the criticisms to the indicator is that it cannot be used in settings where auxiliary information is available only at the population level. In this paper, we propose a new method for estimating response propensities that does not need auxiliary information for non-respondents to the survey and is based on population auxiliary information. These population-based response propensities can then be used to develop R-indicators that employ population contingency tables or population frequency counts. We discuss the statistical properties of the indicators, and evaluate their performance using an evaluation study based on real census data and an application from the Dutch Health Survey.

# 1. Introduction

Nonresponse bias in surveys is of increasing concern with declining response rates and tighter budgets. National Statistics Institutes (NSIs) charged with conducting national surveys to convey the state of their country's economic, social and demographic characteristics are facing increasing challenges in maintaining the quality of their survey response. In this paper, we focus on one particular survey conducted since 1998 by Statistics Netherlands, The Dutch Health Survey which up until 2010 was a face-to-face survey. In 2010, online data collection was added as a sequential mode before the face-to-face interviews. The response rates have gradually declined from values close to 70% to values around 60%. Other NSIs and survey organizations have reported declining response rates, particularly when moving to mixed modes of data collection in order to reduce budgets. However, response rates alone are not enough to judge the quality of the survey response, rather is it the contrast between those responding and not responding to the surveys, or the nonresponse bias. Nonresponse bias in the Dutch Health Survey is conjectured for persons with weaker health, certain habits like smoking or less dentist visits, and worse living conditions. Important predictors are age, marital status, income and ethnicity.

A number of indirect measures of nonresponse bias have been developed recently to supplement the traditional response rate. Wagner (2012) provides a taxonomy of such measures. The most prominent are R-indicators (Schouten, Cobben and Bethlehem 2009, Schouten, Shlomo and Skinner 2011) and balance indicators (Särndal 2011, Lundquist and Särndal 2013). The development of these measures comes at a time where there is an increased interest in adapting data collection (Schouten, Calinescu and Luiten 2013, Wagner 2013, Wagner and Hubbard 2014) in which the level of effort targeted at different subgroups as defined by auxiliary variables may be varied over time, possibly through a change of strategy, according to patterns of response (Schouten, et al. 2012, Särndal and Lundquist 2014).

The auxiliary data used in the measures may stem from sampling frame data, administrative data and data about the data collection process, called paradata (Kreuter 2013). Balance indicators and R-indicators are very similar and are often proportional in size. In this paper, we focus on R-indicators. However, much of the discussion and results can easily be translated to balance indicators.

R-indicators presume the availability of auxiliary variables through linked data from sample frames, registers, etc. to the survey sample. This presumption of linked survey samples is in many settings not a valid one and hampers application. While national statistical institutes often have access to government registrations, university and market researchers usually do not. For indicators to become useful for these researchers, they must be based on different forms of auxiliary information. The only form of auxiliary information that is generally accessible are the sets of statistics produced by the national statistical institutes. These institutes disseminate tables on a wide range of population statistics. This paper develops R-indicators that are based completely on such population statistics and that can be computed without any knowledge about the non-respondents. As an example, market research companies compare the response distributions of a fixed, pre-scribed set of auxiliary variables to national statistics, termed the gold standard. The R-indicator estimators proposed here allow for monitoring and evaluating gold standard variables during and after data collection.

R-indicators and their statistical properties, as discussed in Shlomo, Skinner and Schouten (2012) relate to the case where we have linked sample level auxiliary information for non-respondents. To develop R-indicators based on population statistics, we propose a new method for estimating response propensities that does not need auxiliary information for non-respondents to the survey. They will be called population-based response propensities. To our knowledge, there is no record in the literature about models for response propensities that employ population information only. For large samples, our estimation strategy for response propensities resembles taking the inverse of nonresponse adjustment weights (e.g. Särndal and Lundström 2005). However, the adjustment weights account for both nonresponse selection and sampling variation in auxiliary variable distributions, and the interest here is not in sampling variation. In this respect, the current paper is innovative and may be valuable and relevant to other statistical areas as well. In this paper, we concentrate on the use of population-based response propensities in the computation of R-indicators.

The auxiliary information for population-based response propensities is obtained from population tables and population counts. In order to do so, we first propose estimating response propensities based on population values, by replacing sample covariance matrices and sample means by known population covariances and population means. Next, using population-based response propensities, we compute estimates for the R-indicator. We call the resulting indicator a population-based R-indicator, and we call the traditional R-indicator a sample-based R-indicator. We focus on three research questions:

1. How to extend sample-based response propensities and R-indicators to population-based response propensities and R-indicators?
2. What are the statistical properties of population-based R-indicators?
3. Are the population-based R-indicators practicable in real survey settings?

In Section 2, we propose a new method for estimating population-based response propensities. In Section 3, we briefly review the definitions and methodology behind R-indicators and then consider their estimation in the population-based setting. In Section 4, we present an evaluation study that is based on drawing samples from real Census data under realistic assumptions about non-response in social surveys and evaluate the properties of the population–based R-indicators. In Section 5, we demonstrate the proposed R-indicators on an application from the Dutch Health Survey of the Netherlands. In Section 6, we end with a discussion and present some caveats related to the proposed indicators and future work.

# 2. Population-based response propensities

## 2.1 General notation

We suppose that a sample survey is undertaken, where a sample $s$ is selected from a finite population $U$. The sizes of $s$ and $U$ are denoted $n$ and $N$, respectively. The units in $U$ are labelled $i = 1,2, \dots , N$. The sample is assumed to be drawn by a probability sampling design $p(.)$, where the sample $s$ is selected with probability $p(s)$. The first order inclusion probability of unit $i$ is denoted $\pi_i$ and $d_i = \pi_i^{-1}$ is the design weight. The evaluation study is according to simple random sampling without replacement. Although large-scale national surveys may use more complex two-stage designs, they are generally planned so that all survey units have an equal inclusion probability similar to simple random sampling. We also provide theoretical expressions under the more general complex survey designs.

We suppose that the survey is subject to unit non-response. The set of responding units is denoted $r$, so $r \subset s \subset U$. We denote summation over the respondents, sample and population by $\Sigma_r$, $\Sigma_s$ and $\Sigma_U$, respectively. Let $r_i$ be the response indicator variable so that $r_i = 1$ if unit $i$ responds and $r_i = 0$, otherwise. Hence, $\{i \in s; r_i = 1\}$. We shall suppose that the typical target of inference is a population mean

$$\bar{Y} = N^{-1} \sum_U y_i$$

of a survey variable, taking value for $y_i$ unit $i$.

We suppose that the data available for estimation purposes consists first of the values $\{y_i; i \in r\}$ of the survey variable, observed only for respondents. Secondly, we suppose that information is available on the values $x_i = (x_{1,i}, x_{2,i}, \dots , x_{K,i})^T$ of a vector of auxiliary variables X. We shall usually suppose each $x_{k,i}$ is a binary indicator variable, where $x_i$ represents one or more categorical variables, since this will be the case in the applications we consider, but $x_i$ our presentation allows for general $x_{k,i}$ values. We assume that values of are observed for all respondents so that $\{y_i, x_i; i \in r\}$ is observed.

We distinguish two settings. One in which $x_i$ is known for all sample units, i.e. for both respondents and non-respondents, and one in which $x_i$ is known only at the aggregate level, i.e. the population total $\sum_U x_i$ and/or the population cross-products $\sum_U x_i x_i^T$. We refer to the two types of information as sample-based auxiliary information and aggregate population-based auxiliary information. The first setting is relevant if the variables making up X are available on a register. However, as outlined in Section 1, in many countries and surveys the availability of auxiliary information on non-respondents may be limited and the second setting may be more useful.

## 2.2 Definition of response propensities

The theory of propensity scores was introduced by Rosenbaum and Rubin (1983) and discussed in the context of survey nonresponse by Little (1986; 1988). Response propensities are defined as the conditional expectation of the response indicator variable $r_i$ given the values of specified variables and survey conditions

$$\rho_X(\boldsymbol{x}_i) = E_m(r_i \mid \boldsymbol{x}_i)$$

where the vector of auxiliary variables is defined as in Section 2.1. For simplicity, we shall write $\rho_i = \rho_X(x_i)$ and hence denote the response propensity just by $\rho_i$. $E_m(.)$ denotes expectation with respect to the model underlying the response mechanism. A detailed discussion of response propensities and their properties is presented in Shlomo et al. (2012). They argue that it is desirable to select auxiliary variables constituting $x_i$ in such a way that the missing at random assumption, denoted MAR (Little and Rubin, 2002), holds as closely as possible.

## 2.3 Estimation of response propensities using population-level information

In case of sample-based auxiliary information, it is possible to estimate response propensities for all sampled units by means of regression models

$$g(\rho_i) = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

where $g(.)$ is a link function, $r_i$ is the dependent variable, and $x_i$ is a vector of explanatory variables. Generally, the response propensities are modelled by generalized linear models. Shlomo et al. (2012) use a logistic link function.

In the population-based setting, it is convenient to consider the identity link function. The identity link function is a good approximation to the more widely used logistic link function when response rates are mid-range, between 20% and 80%, which is the typical response rate obtained in national and other surveys. We demonstrate this fact in the evaluation study presented in Section 4 where three ranges of response rates are investigated: low, medium and high. The identity link function also forms the basis for other representativeness indicators in the literature, such as the imbalance and distance indicators proposed by Särndal (2011).

Under the identity link function we assume that the true response propensities satisfy the 'linear probability model'

$$\rho_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad i \in U \tag{1}$$

The linear probability model in (1) can be estimated by weighted least squares, where $d_i$ is the design weight. The implied estimator of $\rho_i$ is given by

$$\hat{\rho}_i^{OLS} = \boldsymbol{x}_i^T \left(\sum_s d_i \boldsymbol{x}_i \boldsymbol{x}_i^T\right)^{-1} \sum_s d_i \boldsymbol{x}_i r_i, \quad i \in s \tag{2}$$

In the case of population-based auxiliary information, we first note that $\sum_s d_i x_i$ and $\sum_s d_i x_i x_i^T$ are unbiased for $\sum_U x_i$ and $\sum_U x_i x_i^T$, respectively, and that in large samples we may expect that $\sum_s d_i x_i \approx \sum_U x_i$ and $\sum_s d_i x_i x_i^T \approx \sum_U x_i x_i^T$. It follows from (2) that, in the population-based setting, we may approximate $\hat{\rho}_i^{OLS}$ by

$$\tilde{\rho}_{i,T1} = \boldsymbol{x}_i^T \boldsymbol{T}_1^{-1} \sum_r d_k \boldsymbol{x}_k, \quad i \in r \tag{3}$$

where

$$\boldsymbol{T}_1 = \sum_U \boldsymbol{x}_j \boldsymbol{x}_j^T.$$

Notice that $\tilde{\rho}_{i,T1}$ is computed only for responding units.

The estimator in (3) requires knowledge of the population sums of squares and cross-products $\sum_U x_i x_i^T$ of the elements of $x_i$. However, the cross-products might be unknown. In that case, we can estimate $\sum_s d_i x_i x_i^T$ in (2) by rewriting

$$\sum_s d_i \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_s d_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s)^T + N \overline{\boldsymbol{x}}_s \overline{\boldsymbol{x}}_s^T, \tag{4}$$

where

$$\overline{\boldsymbol{x}}_s = \sum_s d_i \boldsymbol{x}_i / N.$$

The sample mean $\bar{x}_s$ may be replaced by the population mean $\bar{x}_U$ and the covariance matrix

$$\boldsymbol{S}_{xx} = N^{-1} \sum_s d_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_s)^T \tag{5}$$

may be replaced by

$$\hat{\boldsymbol{S}}_{xx} = \left(\sum_s d_j r_j\right)^{-1} \sum_s d_i r_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_U)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_U)^T. \tag{6}$$

Combining (3), (4) and (6), we obtain the following estimator

$$\tilde{\rho}_{i,T2} = \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \sum_r d_k \boldsymbol{x}_k, \quad i \in r, \tag{7}$$

where

$$\hat{\boldsymbol{T}}_2 = N \hat{\boldsymbol{S}}_{xx} + N \overline{\boldsymbol{x}}_U \overline{\boldsymbol{x}}_U^T.$$

Note that (6) is subject to nonresponse bias itself. We may refine (6) by constructing an iterative algorithm that first computes (6) and (7) and then alternates two steps until convergences. The two steps are 1) use estimates (7) as added inverse propensity weights in (6), and 2) re-estimate (7). We leave this to future study.

In the following, we distinguish between two types of aggregated population auxiliary information as denoted by the indices 'T1' in (3) and 'T2' in (7):

TYPE 1: Full aggregate population-based auxiliary information: the population cross products are available, i.e. $\sum_U x_i x_i^T$ or $\sum_U (x_i - \bar{x}_U)(x_i - \bar{x}_U)^T$;

TYPE 2: Marginal aggregate population-based auxiliary information: only the population marginal counts are available, i.e. $\sum_U x_i$.

The first type implies that we have available all two by two tables, e.g. age times gender, age times marital status and gender times marital status. This information might be available to a national statistical institute who has access to population registers or detailed population demographics and wish to use population-based information to monitor data collection due to a lack of sample-based information on the sample frames. The second type is more restrictive as we have only frequency counts, e.g. age, gender, marital status, without any knowledge about the interactions. This information would be routinely available through websites of national statistical institutes and therefore can be used by marketing and other data collection agencies to monitor their data collection.

# 3. Estimation of R-indicators based on population totals

In this section, first we briefly review the definition and concepts of R-indicators, and their estimation based on sample-level information. Details can be found in Shlomo et al. (2012). Next, applying the theory introduced in Section 2.3, we adapt the sample-based R-indicator to the case where auxiliary information is obtained from population tables and population counts. Further, we investigate the statistical properties of this estimator.

## 3.1 R-indicators

Schouten, et al. (2009) introduce the concept of representative response. A response to a survey is said to be representative with respect to $X$ when response propensities are constant for $X$. The overall measure of representative response is the R-indicator. The R-indicator associated with a set of population response propensities $\{\rho(X_i): i \in U\}$, is defined as

$$R_\rho(X) = 1 - 2S_\rho(X), \qquad (8)$$

where $S_\rho(X)$ denotes the standard deviation of the individual response propensities

$$S_\rho^2(X) = \frac{1}{N-1}\sum_U(\rho(X_i) - \bar{\rho}_U)^2 = \frac{N}{N-1}\left\{\frac{1}{N}\sum_U \rho^2(X_i) - \left[\frac{1}{N}\sum_U \rho(X_i)\right]^2\right\}. \quad (9)$$

The reference to the auxiliary vector $X$ is crucial; for each vector the R-indicator will, generally, be different. In the following, we will, however, often omit the reference to $X$ in order to avoid complex notation.

The R-indicator takes values on the interval $[1 - \sqrt{\frac{N}{N-1}}, 1]$ with the upper value 1 indicating the most representative response, where the $\rho_i$'s display no variation, and the lower value $1 - \sqrt{\frac{N}{N-1}}$ (which is close to 0 for large surveys) indicating the least representative response, where the $\rho_i$'s display maximum variation.

An important related measure of representativeness is the coefficient of variation of the response propensities

$$CV_\rho(X) = \frac{S_\rho(X)}{\bar{\rho}_U}. \quad (10)$$

This is a relevant measure when considering population means or totals as parameters of interest. In those cases, it may be used instead of the R-indicator. The coefficient of variation bounds the absolute non-response bias of standardized unadjusted response means for an arbitrary linear combination of the auxiliary variables in $X$, where the standardization is by the standard deviation of the linear combination. It  Schouten et al. (2016) also used the coefficient of variation to assess 'worst case' non-response bias intervals for standard non-response adjusted post-survey estimators (such as generalized regression estimator (GREG) (Deville and Särndal,  1992) and inverse propensity weighting (IPW) (Little, 1988).

It is not true that (10) also bounds the absolute bias of arbitrary variables that are not linear combinations of the auxiliary variables in $X$. However, (10) is used from the viewpoint that larger bias on auxiliary variables is a sign of larger risk of bias on other variables.

## 3.2 Sample-based R-indicators

In case of sample-based auxiliary information, it is possible to estimate response propensities for all sampled units. In the following, let $\hat{\rho}_i$ be an estimator for $\rho_i$. The sample-based estimator for the R-indicator is

$$\hat{R}_{\hat{\rho}}(X) = 1 - 2\hat{S}_{\hat{\rho}}(X), \quad (11)$$

where $\hat{s}_{\hat{\rho}}^2(X)$ is the design-weighted sample variance of the estimated response propensities computed using the first expression in (9)

$$\hat{S}_{\hat{\rho}}^2 = \frac{1}{N-1}\sum_s d_i\left(\hat{\rho}_i - \hat{\bar{\rho}}_U\right)^2,$$

where

$$\hat{\hat{\rho}}_U = \left(\sum_s d_i \hat{\rho}_i\right) / N \; .$$

The sample-based R-indicator defined by (11) is a statistic with a certain precision and bias. Shlomo et al. (2012) discuss bias adjustments and confidence intervals for $\hat{R}_{\hat{\rho}}$. These are available in SAS and R code at www.risq-project.eu, and a manual is provided by De Heij, Schouten and Shlomo (2015). We return to the statistical properties in section 3.4.

## 3.3 Population-based R-indicators

We demonstrate in Section 4 that the R-indicators depend only mildly on the type of link function when estimating response propensities where response rates are not in the tails, i.e. very high or very low. Furthermore, we obtain similar estimation of R-indicators when population-based response propensities are estimated according to the Type 1 or Type 2 types of information.

In the population-based setting, an estimator for the R-indicator is then

$$\tilde{R}_{\tilde{\rho}}(X) = 1 - 2\tilde{S}_{\tilde{\rho}}(X), \tag{12}$$

where

$$\tilde{S}^2_{\tilde{\rho}} = \frac{N}{N-1}\left\{ \frac{1}{N}\sum_r d_i \tilde{\rho}_i - \left(\frac{1}{N}\sum_r d_i\right)^2 \right\}, \tag{13}$$

and $\tilde{\rho}_i$ denotes either response propensities computed under Type 1 information ($\tilde{\rho}_{i,T1}$) or response propensities estimated under Type 2 information ($\tilde{\rho}_{i,T2}$).

The estimation of the R-indicator is based on the second expression for $S^2_\rho$ in (9). This choice indeed makes the estimator $\tilde{S}^2_{\tilde{\rho}}$ linear in $\tilde{\rho}_i$, which provides an advantage for bias computations as described in Section 3.4. We have compared estimators based on the two expressions (results not included in this paper) and found very small differences for the scenarios that we explored (see the evaluation study in section 4).

Furthermore, we use propensity-weighting by $\tilde{\rho}_i^{-1}$ to adjust for non-response bias. As for standard non-response weighting, the validity of this correction depends on the validity of the estimates $\tilde{\rho}_i$.

We remark that any adjustment technique for non-response can be applied to construct estimators for $R_\rho$, e.g. calibration estimators such as linear or multiplicative weighting (Särndal and Lundström 2005) or weighting class adjustments (Little 1986). It is generally known that propensity weighting may lead to larger standard errors. It may, therefore, be more efficient to use parsimonious models to estimate the R-indicator. For instance, this can be done by stratifying on response propensity classes. However, we did not explore such estimators, and restricted ourselves to the propensity-weighted estimator (12). This is a topic for future research.

The estimation of the coefficient of variation (10) in the population-based setting is straightforward

$$CV_{\tilde{\rho}}(X) = \frac{\tilde{S}_{\tilde{\rho}}(X)}{\tilde{\bar{\rho}}_U}$$

where

$$\tilde{\bar{\rho}}_U = \sum_r d_i \,/\, N .$$

Despite being straightforward estimators, the population-based R-indicators based on (3) and (7) are problematic. Their standard errors and biases increase with higher response rates. We will demonstrate this tendency in the evaluation study in Section 4.2. Clearly, more respondents should provide smaller standard errors and create less bias since the auxiliary variables will not vary as much on the remaining non-response. The reason that (3) and (7) have these properties, is that they are natural but naïve estimators that ignore the sampling which causes sample covariances in the denominator of the estimated response propensities to vary along with the numerator. By 'plugging' in a fixed population covariance in the denominator, there is no variation arising from sampling.

One way to moderate this effect would be to use a composite estimator, i.e. to employ a linear combination of the estimated propensity and the response rate,

$$\tilde{\rho}_{i,T1}^{C} = (1-\lambda)\tilde{\rho}_{i,T1} + \lambda\tilde{\bar{\rho}}_U \tag{14}$$

and similarly for Type 2. The composite estimate in (14) is similar to a 'shrinkage' estimator, e.g. Copas (1983 and 1993), for the variance of the response propensities $\tilde{S}_{\tilde{\rho}}^2$ given by (13). In that case, the optimal $\lambda$ is usually chosen by minimizing the MSE by solving the derivative of the MSE with respect to $\lambda$. We return to the choice of $\lambda$ in Section 3.4 and note here that given the observed bias and variance properties, $\lambda$ should be an increasing function of the response rate and should converge to 1 with higher response rates. Such a $\lambda$ will draw potential estimated response propensities greater than 1 due to the use of the linear link function under high response rates to be closer to 1.

We explored several other possible alternatives to the composite estimator in (14), including a Hájek-type adjustment, a composite estimator of the population covariance matrix and the response covariance matrix of the $x_i$, and response propensities truncated to the interval [0,1] for high response rates. However, all gave worse results than the composite estimator in the propensities as shown in (14). Furthermore, approximate bias adjustments of the R-indicators using (14) can be easily constructed from those using (3) and (7).

A promising alternative may be to adopt an EM-algorithm approach in which the missing auxiliary variables for nonrespondents are imputed. Such an approach is, however, very different in nature and we leave this to future research.

## 3.4 Bias and standard error of the population-based R-indicator

Shlomo et al. (2012) derive analytic approximations for the bias and standard errors of the sample-based estimate for the R-indicator (11) under both the sampling and nonresponse random mechanisms. The bias in this estimator arises mostly from 'plugging in' estimated response propensities in the sample variances. This source of bias is referred to as small sample bias. A much smaller and usually negligible contribution to the bias originates from using sample means rather than population means. Even if the response is representative, i.e. has equal response propensities, some variation in estimated response propensities is found. The bias is inversely proportional to the sample size. The larger the sample, the smaller the bias. Schouten, et al. (2009) investigate the bias for different sample sizes. From their analyses, it follows that the bias is relatively small for typical sample sizes used in large-scale surveys in comparison to the standard error of the R-indicators. Also, the bias adjustment is successful in removing the bias.

For the estimated population-based R-indicators, we expect that statistical properties will be quite different from their sample-based counterparts. As these estimators use less information, the standard errors will be larger. The bias of the population-based estimators may also be larger since in addition to the bias that was evident for small sample sizes in the sample-based estimators, the population-based estimators will likely have bias arising from the estimation of the sample means and covariances and from the restriction to (propensity-weighted) response means.

To reduce the bias of the population-based estimators, we propose to adjust the bias of $\tilde{S}^2_{\tilde{\rho}_{T1}}$ and $\tilde{S}^2_{\tilde{\rho}_{T2}}$, respectively. This leads to the adjusted version of the estimator for the R-indicator under Type 1 information

$$\tilde{R}^{ADJ}_{\tilde{\rho}_{T1}} = 1 - 2\left[\tilde{S}^2_{\tilde{\rho}_{T1}} - \tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)\right]^{1/2}. \tag{15}$$

Appendix A derives the general expression for $\tilde{B}_{\tilde{\rho}_{T1}}(\tilde{S}^2_{\tilde{\rho}_{T1}})$ under both simple random sampling and a more general expression under complex sampling. From Appendix A, the response-set based estimator for the bias under simple random sampling is

$$\tilde{B}^{SRS}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1 - \frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i + \frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n} - \frac{n_r}{n^2}\right], \tag{16}$$

where $n_r$ denotes the size of the response set $r$.

In the case of Type 2 information, the adjusted version of the estimator for the R-indicator is as (15) with the Type 2 terms replacing the Type 1 information.

Appendix B derives the general expression for the bias of $\tilde{S}^2_{\tilde{\rho}_{T2}}$, $\tilde{B}_{\tilde{\rho}_{T2}}(\tilde{S}^2_{\tilde{\rho}_{T2}})$, under simple random sampling and the more general case of complex sampling. From Appendix B, the response-set based estimator for the bias under simple random sampling is

$$\tilde{B}^{SRS}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right)=\frac{N}{N-1}\left\{\frac{1}{n_r^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{F}}\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{t}}-\frac{N}{nn_r}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^T\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{t}}\right.$$
$$\left.+\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\boldsymbol{x}_i+\frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T2}-\left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T2}}}{n}-\frac{n_r}{n^2}\right\},$$

where

$$\hat{\boldsymbol{F}}=Nn^{-1}\sum\nolimits_r \boldsymbol{z}_k\boldsymbol{z}_k^T$$
$$\hat{\boldsymbol{t}}=Nn^{-1}\sum\nolimits_r \boldsymbol{x}_k$$
$$\boldsymbol{z}_i=\left(\boldsymbol{x}_i-\overline{\boldsymbol{x}}_U\right).$$

Turning to the composite estimator, it is straightforward to show that combining (13) and (14) leads to

$$\tilde{S}^2_{\tilde{\rho}^C_{T1}}=(1-\lambda)\,\tilde{S}^2_{\tilde{\rho}_{T1}}, \tag{17}$$

and its bias equals

$$B\left(\tilde{S}^2_{\tilde{\rho}^C_{T1}}\right)=(1-\lambda)\,B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)-\lambda S^2_{\rho}. \tag{18}$$

A response-set based estimator for $B(\tilde{S}^2_{\tilde{\rho}^C_{T1}})$ is obtained using the response-set based estimator developed for $B(\tilde{S}^2_{\tilde{\rho}_{T1}})$. For the Type 1 estimator and under simple random sampling

$$\tilde{B}^{SRS}_{\tilde{\rho}^C_{T1}}\left(\tilde{S}^2_{\tilde{\rho}^C_{T1}}\right)=(1-\lambda)\,\tilde{B}^{SRS}_{\tilde{\rho}^C_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)-\lambda\tilde{S}^2_{\tilde{\rho}^C_{T1}}$$
$$=(1-\lambda)\frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}^C_{i,T1}\right\}\mathbf{x}_i^T\mathbf{T}_1^{-1}\mathbf{x}_i+\frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}^C_{i,T1}-\left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}^C_{T1}}}{n}-\frac{n_r}{n^2}\right]-\lambda\tilde{S}^2_{\tilde{\rho}^C_{T1}}. \tag{19}$$

The same approach applies for Type 2 estimator.

The variance of (17) is equal to

$$\mathrm{V}\left(\tilde{S}^2_{\tilde{\rho}^C_{T_1}}\right)=(1-\lambda)^2\mathrm{V}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right). \tag{20}$$

To estimate the variance of $\tilde{S}^2_{\tilde{\rho}_{T_1}}$ defined in (13), $\mathrm{V}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$, which is needed to estimate the variance of $\tilde{R}^{ADJ}_{\tilde{\rho}_{T_1}}$ in (15) as well as the variance of the composite estimator in (20), we use resampling methods. More specifically, we employ bootstrap methods (see: Efron and Tibshirani 1993 and Booth et al. 1994 and Wolter 2007 for the use of bootstrapping methods for finite populations) and assess their performance in the evaluation study in Section 4.

We return now to the  choice of $\lambda$ for the composite estimator in (14), the optimal $\lambda$ can be derived by combining (18) and (20), and then take derivatives. Letting $B$ and $V$ denote $B\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$ and $\mathrm{V}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$, respectively, it follows that the optimal $\lambda$ is

$$\lambda_{\mathrm{opt}} = \frac{B\left(B+S_\rho^2\right)+V}{\left(B+S_\rho^2\right)^2+V}. \tag{21}$$

In order to estimate a value for $\lambda_{\mathrm{opt}}$, $B, V$ and $S_\rho^2$ need to be estimated. Under Type 1 information and simple random sampling, we propose to estimate $B$ by $B^{SRS}_{\tilde{\rho}_{T_1}}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$ as in (16), $S_\rho^2$ by $\tilde{S}^2_{\tilde{\rho}_{T_1}}$, and $V$ by the bootstrap variance estimator of $\tilde{S}^2_{\tilde{\rho}_{T_1}}$. This leads to the population-based Type 1 estimator for $\lambda_{\mathrm{opt}}$, denoted $\tilde{\lambda}_{opt,T1}$, and the population-based composite propensities

$$\tilde{\rho}^{PC}_{i,T1} = \left(1-\tilde{\lambda}_{opt,T1}\right)\tilde{\rho}_{i,T1} + \tilde{\lambda}_{opt,T1}\bar{\tilde{\rho}}_U .$$

The corresponding population-based R-indicator is then computed as in (12) and its bias-adjusted version as in (15), where the bias adjustment is given by (19).

We propose to estimate the variance of the population-based composite estimator, $\mathrm{V}\left(\tilde{S}^2_{\tilde{\rho}^{PC}_{T1}}\right)$, by linearization

$$\hat{V}\left(\tilde{S}^2_{\tilde{\rho}^{PC}_{T1}}\right) = \frac{\tilde{V}^{BT}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)\left(1-\tilde{\lambda}_{opt,T1}\right)^2}{\tilde{S}^2_{\tilde{\rho}_{T_1}}},$$

where $\tilde{V}^{BT}\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$ is the bootstrap variance estimator for $V\left(\tilde{S}^2_{\tilde{\rho}_{T_1}}\right)$.

The same approach applies for Type 2 information.

# 4. Evaluation study

In this section, we carry out an evaluation study on real census data from the 1995 Israel Census Sample to assess the sampling properties of the estimation procedures introduced in Section 3.

The aim of the evaluation is two-fold: a) study sampling properties of the unadjusted and bias adjusted population-based R-indicators, comparing them to those of their sample-based counterpart and assessing the effect of sample size, number of auxiliary variables in the model, and response rate; b) investigate the performance of the bootstrap estimator for estimating the variance of the population-based R-indicator.

## 4.1 Data and design of the evaluation study

The 1995 20% Israel Census Sample contains 753,711 individuals aged 15 and over in 322,411 households. The sample design is similar to a standard household survey carried out at national statistical institutes. The sample units are households and all persons over the age of 15 in the sampled households are interviewed. Typically a proxy questionnaire is used and therefore there is no individual non-response within the household. In this study, we assume that every household has an equal probability to be included in the sample which is standard practice in large-scale social surveys. The evaluation uses data at the household level.

We carried out a two-step design to define response propensities in the population (census) data. This procedure ensures that we have a known model generating the response propensities. Moreover, in order to explore the effect of varying response rates and the number of auxiliary variables in the model on the performance of the estimators, we considered six scenarios.

a. First, probabilities of response were defined according to variables: Type of locality (3 categories), number of persons in household (1,2,3,4,5,6+), children in the household indicator (yes, no), region (7 categories), and density (3 categories). These variables define groups that are known to have differential response rates for social surveys in practice. To study the effect of response rates on the performance of the estimators, probabilities of response p were defined according to $p = p_1 p_2 p_3 p_4 p_5 + \alpha$ with four choices $\alpha = 0.15$ (RR1), $\alpha = 0.55$ (RR2), and $\alpha = 0.75$ (RR3), where the probabilities $p_1$, $p_2$, $p_3$, $p_4$ and $p_5$ are given in Table 4.1. We generate three response indicator variables using the Bernoulli distribution for each of the response scenarios defined under RR1, RR2, and RR3.

b. For each of the response scenarios from step (A), we use the response indicator as the dependent variable and fit both a linear and a logistic regression model to the population to predict 'true' response propensities for our evaluation study under both link functions. Two different models were considered for prediction of 'true' response propensities. In Model 1, independent variables are exactly the explanatory variables used in step A for the definition of response probabilities (child indicator, number of persons in the household, region, type of locality, density). In Model 2, independent variables are type of locality, number of persons in household, child indicator. Notice that we use the same response indicator variables to fit the two models. This allows to isolate the effect of the model, excluding differences due to random variability in the response indicator. In both models, the auxiliary variables are included as main effects only, i.e. excluding interactions terms. In the evaluation, we compare to population R-indicators computed using main effect models as well in order to avoid differences due to model misspecification.

Response rates for the variables defining probabilities as well as the overall response rates and population values of the R-indicator under the two models are shown in Table 4.1. For comparison purposes, we report population values of the R-indicator based on both linear and logistic regression models where the response rates range between 25.1% and 35.1% under RR1, between 64.7% and 75.4% under RR2 and between 84.7% and 94.6% under RR3. RR2 represents the type of response rate seen in large-scale national social surveys. As can be seen in Table 4.1, there is little difference in the population values of the R-indicators based on the linear and logistic link function for RR1 and RR2 and a slight difference for RR3 where response rates are in the upper tail of the distribution.

Table 4.1: Probabilities of response and percentage response generated in the evaluation population dataset according to auxiliary variables

| Variable | Category | Probabilities | RR1 | RR2 | RR3 |
|---|---|---|---|---|---|
| Children in Household | None | $p_1$ =0.6 | 25.7 | 65.6 | 85.7 |
| | 1+ | $p_1$ =0.8 | 35.1 | 75.4 | 94.6 |
| Number of Persons in Household | 1-2 | $p_2$ =0.5 | 24.6 | 64.5 | 84.7 |
| | 3-5 | $p_2$ =0.8 | 32.9 | 72.8 | 92.5 |
| | 6+ | $p_2$ =0.7 | 29.9 | 70.3 | 90.0 |
| Type of Locality | Type 1 | $p_3$ =0.6 | 25.1 | 64.9 | 85.0 |
| | Type 2 | $p_3$ =0.7 | 28.3 | 68.5 | 88.4 |
| | Type 3 | $p_3$ =0.8 | 31.5 | 71.7 | 91.2 |
| | Type 4 | $p_3$ =0.75 | 28.9 | 69.2 | 88.9 |
| Region | 1 | $p_4$ =0.6 | 25.1 | 65.1 | 84.7 |
| | 2 | $p_4$ =0.8 | 31.2 | 71.5 | 91.0 |
| | 3 | $p_4$ =0.7 | 28.1 | 67.6 | 87.8 |
| | 4 | $p_4$ =0.6 | 26.7 | 66.5 | 86.4 |
| | 5 | $p_4$ =0.6 | 24.8 | 64.7 | 84.9 |
| | 6 | $p_4$ =0.7 | 27.6 | 67.8 | 88.0 |
| | 7 | $p_4$ =0.8 | 30.3 | 70.4 | 90.9 |
| Density | <=1.5 | $p_5$ =0.6 | 26.1 | 66.0 | 86.2 |
| | 1.5-3.0 | $p_5$ =0.8 | 28.9 | 68.9 | 88.8 |
| | >3 | $p_5$ =0.7 | 24.7 | 64.7 | 84.7 |
| Overall response rate | | | 27.1 | 67.0 | 87.0 |
| Population R-indicator (logistic) | Model 1 | | 0.9031 | 0.9005 | 0.9063 |
| | Model 2 | | 0.9103 | 0.9074 | 0.9137 |
| Population R-indicator (linear) | Model 1 | | 0.9033 | 0.9006 | 0.9076 |
| | Model 2 | | 0.9104 | 0.9074 | 0.9145 |

When using Model 2, the R-indicator is always around 0.007 points greater than the corresponding value under Model 1. This is due to the fact that Model 2 for estimating the response propensities is miss-specified in this case. There are less auxiliary variables and hence less variation in the estimated response propensities which leads to a higher R-indicator. As a consequence we obtain a slightly higher R-indicator for Model 2 as some of the variation is not captured. For this reason, it is important to report R-indicators together with the auxiliary information used to calculate them and in addition to use covariates that correlate to the survey variables (Schouten, et al. 2012).

For each response scenario, five hundred samples were drawn from the population under simple random sampling (SRS) at three different sampling rates 1%, 2%, and 4%. For each sample drawn, a sample response indicator was generated from the 'true' population response probability based on the logistic link function. This determines the response set r. Response propensities and R-indicators were then estimated from each sample for both sample and population-based auxiliary variables. Response propensities are estimated in the sample using the 'true' model (either Model 1 or Model 2, depending on the scenario).

In order to estimate the variance of population-based estimators, we employ a non-parametric bootstrap algorithm. From each response set, we drew B=500 bootstrap samples using simple random sampling (SRS) with replacement. Subsequently, non-response was generated in the bootstrap sample by copying the 0-1 sample response indicator values. A replicate of the estimator was computed over each bootstrap sample.

## 4.2 Results

We contrast the sample-based R-indicators (under both link functions) with the population-based R-indicators. In the evaluation, we also investigate the performance of the population-based composite estimator (PC) as shown in (14). Figures 4.1 to 4.3 present box plots comparing the estimators and their bias adjusted versions under Model 1, and different response rates RR1, RR2 and RR3, respectively.

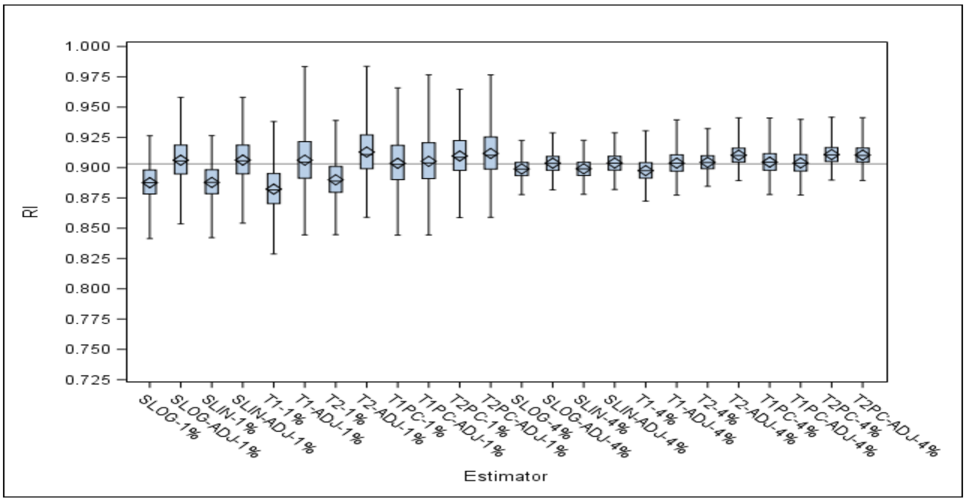


Figure 4.1: Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR1. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

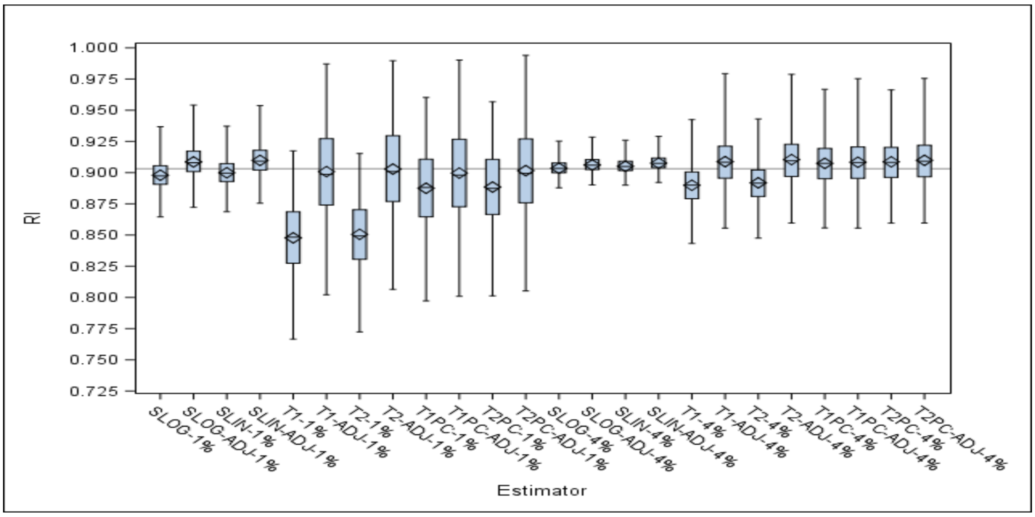


Figure 4.2: Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR2. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

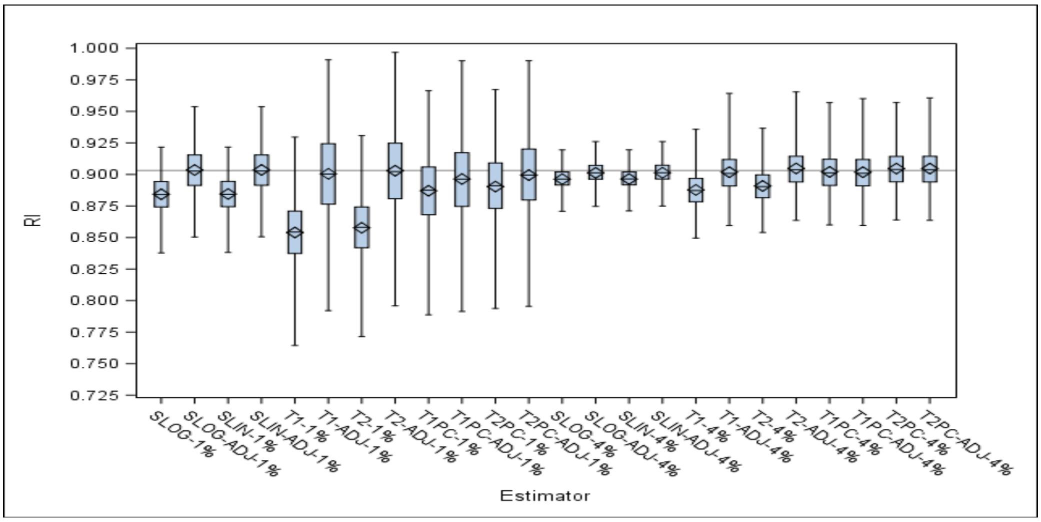


Figure 4.3: Boxplots for 500 estimated R-indicators for 1% and 4% samples for Model 1 and RR3. (SLOG) denotes the logistic sample-based R-indicator, (SLIN) the linear sample-based R-indicator, (T1) the Type 1 population-based R-indicator, (T2) the Type 2 population-based R-indicator, and (T1PC) and (T2PC) the Type 1 and Type 2 population-based composite estimators. ADJ refers to the corresponding bias-adjusted estimators.

Figures 4.1 to 4.3 tell us that the gains from the bias adjustments are evident for Type 1 and Type 2 R-indicators. Standard errors for RR3 are much larger than for RR1 under the same sampling rates. The variability of the bias-adjusted estimator increases and it is larger for smaller sample sizes.

Table 4.2 presents results of the evaluation study for each response rate scenario, type of model and each sampling rate. For each estimator, Table 4.2 shows: a) the percentage Relative Bias (%RB) calculated as

$$100\left\{\left[\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j}-R_{\rho}\right)/R_{\rho}\right]/500\right\},$$

where $\hat{R}_{\hat{\rho}j}$ is the value of the estimator computed for the j-th sample and $R_{\rho}$ is the true R-indicator based on the linear regression model (from Table 4.1), and similarly for $\tilde{R}_{\tilde{\rho}_{T1}}$, $\tilde{R}_{\tilde{\rho}_{T2}}$, and the composite estimator, and b) the Relative Root Mean Square Error (RRMSE) calculated as

$$100\left\{R_{\rho}^{-1}\sqrt{\sum_{j=1}^{500}\left(\hat{R}_{\hat{\rho}j}-R_{\rho}\right)^{2}/500}\right\}.$$

Table 4.2 shows that differences between the sample-based estimators computed using the linear and the logistic link functions are very small in general, except when the response rates get very close to 1 (RR3).

For sample-based and population-based Type 1 and Type 2 estimators there is a general downward bias in the unadjusted R-indicators and this tends to decrease as the sample size increases for both Models 1 and 2. This is as expected. Sampling error tends to lead to overestimation of the variability of the estimated response propensities and this leads to underestimation of the R-indicator. The degree of underestimation is generally larger for population-based estimators than for the sample-based estimators, especially for the higher response rate The variation of response propensities is larger in this case than the variation under sample-based auxiliary variables. In addition, the RRMSE of the estimators decrease as sample size increases and is generally larger for population-based estimators. Thus, the population-based R-indicators are in general less accurate than their sample-based counterparts and allow for weaker conclusions regarding the nature of response.

In general, the unadjusted population-based composite estimators have a better performance than the corresponding unadjusted population-based estimators, both in terms of %RB and RRMSE, especially for higher response rates. They still show some degree of overestimation under the correct Model 1 for low response rates and underestimation for high response rates. However, for Model 2 we see overestimation.

We now turn to the bias-adjusted estimated R-indicators in Table 4.2. For Type 1, the bias adjustment is able to remove the bias. The analytical bias adjustment for Type 1 population-based estimator works well and generally outperforms the analytical bias adjustment for Type 2 population-based estimates. It seems to pick up most of the bias and provides adjusted estimates that are closer to sample-based R-indicators. The RRMSE for the bias-adjusted estimator is generally similar to the corresponding RRMSE for the unadjusted estimator, meaning that the increase in variability is compensated by the bias reduction. For higher response rates, the adjusted population-based composite estimate reduces the bias and RRMSE of their corresponding population based R-indicators.

In unadjusted form, the Type 2 R-indicator behaves better than the Type 1 R-indicator. This is rather surprising as we seem to be able to have more accurate estimation of the true R-indicator when using less information. The reason for this is that for the Type 1 estimator we do not include any of the sampling variation when we 'plug in' the population covariance matrix, whilst for the Type 2 estimator we use only the marginal information and 'plug in' the response covariance matrix which accounts for more of the sampling variation. After the bias adjustment, the Type 2 estimators have higher %RB (especially for lower response rates) but similar RRMSE. Type 2 bias adjustment performs worse than the bias adjustment for Type 1 and overcompensates for the bias. This result was expected as the Type 2 bias adjustment is based on a linear approximation, while Type 1 bias adjustment is computed exactly.

Regarding increasing response rates, surprisingly, for the population-based unadjusted estimators, we observe a better performance for lower response rates, both in terms of percentage relative bias (%RB) and RRMSE. The RRMSE of RR3 are 2 to 3 times larger than for RR1. Analytical bias adjustments work very well under all response rates, although with higher RRMSEs for higher response rates. This problem is reduced by the use of the composite estimators.

Regarding the effect of the number of variables in the model, a lower %RB and RRMSE are observed under Model 2 for unadjusted population-based estimators compared to Model 1. The composite estimators show in general an opposite pattern. The bias-adjusted versions show similar performance under the two models.

Table 4.2: Properties of the estimated R-indicators for sample and population-based auxiliary variables for 500 samples in the evaluation study.

| Rate | Sample Rate | Estimator | Model 1 | | | | Model 2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Unadjusted | | Adjusted | | Unadjusted | | Adjusted | |
| | | | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE | %RB | RRMSE |
| RR1 | 1% | Sample | -1.73 | 2.39 | 0.32 | 2.01 | -0.77 | 1.88 | 0.34 | 1.96 |
| | | Sample | -1.71 | 2.37 | 0.33 | 2.01 | -0.75 | 1.87 | 0.35 | 1.95 |
| | | Type1 | -2.32 | 3.08 | 0.32 | 2.54 | -1.08 | 2.32 | 0.30 | 2.39 |
| | | Type1-PC | 0.04 | 2.28 | 0.22 | 2.42 | 0.59 | 2.44 | 0.38 | 2.41 |
| | | Type2 | -1.47 | 2.29 | 1.06 | 2.50 | -0.20 | 1.74 | 1.01 | 2.27 |
| | | Type2-PC | 0.71 | 2.11 | 0.94 | 2.34 | 1.19 | 2.32 | 1.05 | 2.28 |
| | 2% | Sample | -0.90 | 1.53 | 0.14 | 1.36 | -0.41 | 1.30 | 0.14 | 1.31 |
| | | Sample | -0.89 | 1.51 | 0.16 | 1.36 | -0.40 | 1.29 | 0.15 | 1.31 |
| | | Type1 | -1.24 | 1.89 | 0.12 | 1.61 | -0.51 | 1.57 | 0.17 | 1.59 |
| | | Type1-PC | 0.04 | 1.56 | 0.10 | 1.59 | 0.38 | 1.68 | 0.21 | 1.62 |
| | | Type2 | -0.45 | 1.30 | 0.84 | 1.64 | 0.26 | 1.31 | 0.86 | 1.63 |
| | | Type2-PC | 0.72 | 1.53 | 0.82 | 1.61 | 1.02 | 1.75 | 0.89 | 1.66 |
| | 4% | Sample | -0.48 | 1.00 | 0.05 | 0.93 | -0.27 | 0.90 | 0.00 | 0.88 |
| | | Sample | -0.46 | 0.99 | 0.06 | 0.92 | -0.26 | 0.89 | 0.01 | 0.88 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Type1 | -0.63 | 1.23 | 0.05 | 1.12 | -0.34 | 1.13 | -0.01 | 1.11 |
| | | Type1-PC | 0.15 | 1.14 | 0.07 | 1.12 | 0.18 | 1.18 | 0.01 | 1.13 |
| | | Type2 | 0.12 | 0.92 | 0.78 | 1.25 | 0.40 | 1.01 | 0.69 | 1.19 |
| | | Type2-PC | 0.83 | 1.29 | 0.79 | 1.26 | 0.83 | 1.30 | 0.70 | 1.20 |
| RR2 | 1% | Sample | -1.81 | 2.44 | 0.33 | 2.01 | -0.76 | 1.83 | 0.34 | 1.94 |
| | | Sample | -1.79 | 2.42 | 0.34 | 2.01 | -0.75 | 1.82 | 0.35 | 1.94 |
| | | Type1 | -5.17 | 5.95 | -0.01 | 3.95 | -2.45 | 3.77 | 0.25 | 3.43 |
| | | Type1-PC | -1.50 | 3.58 | -0.47 | 3.69 | 0.69 | 3.37 | 0.49 | 3.46 |
| | | Type2 | -4.76 | 5.50 | 0.27 | 3.75 | -1.95 | 3.29 | 0.58 | 3.23 |
| | | Type2-PC | -1.13 | 3.28 | -0.12 | 3.51 | 0.74 | 3.13 | 0.71 | 3.25 |
| | 2% | Sample | -1.00 | 1.59 | 0.08 | 1.37 | -0.40 | 1.29 | 0.14 | 1.30 |
| | | Sample | -0.98 | 1.57 | 0.09 | 1.36 | -0.40 | 1.28 | 0.14 | 1.30 |
| | | Type1 | -2.89 | 3.55 | 0.07 | 2.59 | -1.19 | 2.58 | 0.37 | 2.72 |
| | | Type1-PC | -0.57 | 2.37 | -0.12 | 2.49 | 0.53 | 2.67 | 0.41 | 2.69 |
| | | Type2 | -2.52 | 3.19 | 0.39 | 2.50 | -0.79 | 2.28 | 0.69 | 2.63 |
| | | Type2-PC | -0.26 | 2.19 | 0.19 | 2.37 | 0.81 | 2.58 | 0.71 | 2.60 |
| | 4% | Sample | -0.48 | 0.98 | 0.07 | 0.90 | -0.16 | 0.81 | 0.12 | 0.83 |
| | | Sample | -0.46 | 0.97 | 0.08 | 0.90 | -0.15 | 0.81 | 0.12 | 0.82 |
| | | Type1 | -1.42 | 2.12 | 0.13 | 1.81 | -0.60 | 1.66 | 0.16 | 1.67 |
| | | Type1-PC | 0.16 | 1.77 | 0.14 | 1.80 | 0.37 | 1.76 | 0.20 | 1.69 |
| | | Type2 | -1.07 | 1.82 | 0.46 | 1.78 | -0.25 | 1.47 | 0.47 | 1.63 |
| | | Type2-PC | 0.45 | 1.72 | 0.46 | 1.75 | 0.65 | 1.73 | 0.50 | 1.66 |
| RR3 | 1% | Sample | -1.07 | 1.59 | 0.10 | 1.30 | -0.52 | 1.21 | 0.02 | 1.16 |
| | | Sample | -0.85 | 1.40 | 0.24 | 1.26 | -0.41 | 1.13 | 0.10 | 1.13 |
| | | Type1 | -6.60 | 7.32 | -0.76 | 4.24 | -3.20 | 4.61 | 0.06 | 4.12 |
| | | Type1-PC | -2.22 | 4.15 | -0.88 | 4.16 | -0.28 | 3.70 | 0.09 | 3.92 |
| | | Type2 | -6.29 | 6.99 | -0.53 | 4.08 | -2.85 | 4.25 | 0.27 | 3.95 |
| | | Type2-PC | -2.12 | 3.97 | -0.67 | 4.02 | -0.04 | 3.52 | 0.33 | 3.78 |
| | 2% | Sample | -0.73 | 1.13 | -0.14 | 0.92 | -0.30 | 0.88 | -0.03 | 0.85 |
| | | Sample | -0.54 | 0.98 | 0.01 | 0.87 | -0.20 | 0.82 | 0.06 | 0.82 |
| | | Type1 | -3.70 | 4.31 | 0.12 | 2.93 | -1.74 | 2.98 | 0.20 | 2.86 |
| | | Type1-PC | -0.78 | 2.60 | -0.15 | 2.78 | 0.42 | 2.81 | 0.36 | 2.94 |
| | | Type2 | -3.46 | 4.07 | 0.30 | 2.87 | -1.46 | 2.73 | 0.41 | 2.77 |
| | | Type2-PC | -0.61 | 2.47 | 0.02 | 2.70 | 0.64 | 2.74 | 0.57 | 2.87 |
| | 4% | Sample | -0.46 | 0.77 | -0.16 | 0.66 | -0.18 | 0.57 | -0.05 | 0.55 |
| | | Sample | -0.29 | 0.66 | -0.01 | 0.61 | -0.09 | 0.53 | 0.04 | 0.53 |
| | | Type1 | -1.96 | 2.62 | 0.12 | 2.12 | -0.89 | 1.81 | 0.13 | 1.76 |
| | | Type1-PC | -0.03 | 1.97 | 0.07 | 2.06 | 0.38 | 1.84 | 0.19 | 1.79 |
| | | Type2 | -1.74 | 2.42 | 0.31 | 2.07 | -0.66 | 1.65 | 0.31 | 1.71 |
| | | Type2-PC | 0.11 | 1.89 | 0.25 | 2.00 | 0.56 | 1.81 | 0.38 | 1.75 |

Table 4.3 shows the mean of the estimated $\lambda_{\text{opt}}$ for the composite population-based Type 1 and Type 2 estimators compared to the true value obtained from the population under the two extreme response rate scenarios, RR1 and RR3. It can be seen that the mean estimated $\lambda_{\text{opt}}$ does not deviate greatly from their true values in the evaluation study.

Table 4.3: Mean $\lambda_{opt}$ for population-based auxiliary variables for 500 samples in the evaluation study

| Rate | Sample rate | Model 1 | | | | Model 2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Type 1 | | Type 2 | | Type 1 | | Type 2 | |
| | | True | Pop | True | Pop | True | Pop | True | Pop |
| RR1 | 1% | 0.40 | 0.33 | 0.36 | 0.33 | 0.31 | 0.29 | 0.26 | 0.28 |
| | 2% | 0.25 | 0.21 | 0.22 | 0.21 | 0.19 | 0.22 | 0.15 | 0.19 |
| | 4% | 0.14 | 0.13 | 0.13 | 0.13 | 0.10 | 0.10 | 0.08 | 0.09 |
| RR3 | 1% | 0.68 | 0.44 | 0.67 | 0.44 | 0.57 | 0.51 | 0.55 | 0.48 |
| | 2% | 0.51 | 0.39 | 0.50 | 0.38 | 0.41 | 0.43 | 0.39 | 0.41 |
| | 4% | 0.35 | 0.27 | 0.34 | 0.27 | 0.25 | 0.23 | 0.24 | 0.22 |

Table 4.4: Properties of variance estimators for R-indicators under sample and population-based auxiliary variables for 500 samples.

| Rate | Sampling rate | Estimator | Model 1 | | Model 2 | |
|---|---|---|---|---|---|---|
| | | | %RB | Coverage | %RB | Coverage |
| RR1 | 1% | Sample-based | 1.84 | 0.95 | -5.74 | 0.95 |
| | | Type 1 | 4.35 | 0.95 | 11.12 | 0.96 |
| | | Type 2 | 4.99 | 0.94 | 7.72 | 0.95 |
| | 2% | Sample-based | 1.43 | 0.96 | 1.15 | 0.95 |
| | | Type 1 | 8.62 | 0.96 | 5.31 | 0.95 |
| | | Type 2 | 7.03 | 0.93 | 2.10 | 0.92 |
| | 4% | Sample-based | 7.93 | 0.97 | -4.58 | 0.95 |
| | | Type 1 | 13.23 | 0.96 | 3.42 | 0.95 |
| | | Type 2 | 13.38 | 0.89 | 2.53 | 0.90 |
| RR3 | 1% | Sample-based | -1.05 | 0.95 | -9.48 | 0.92 |
| | | Type 1 | 2.87 | 0.78 | 11.47 | 0.86 |
| | | Type 2 | 4.97 | 0.78 | 10.26 | 0.85 |
| | 2% | Sample-based | -4.34 | 0.94 | -7.96 | 0.94 |
| | | Type 1 | -7.61 | 0.92 | 2.37 | 0.91 |
| | | Type 2 | -8.07 | 0.92 | 1.02 | 0.90 |
| | 4% | Sample-based | 3.31 | 0.94 | -3.54 | 0.95 |
| | | Type 1 | -8.33 | 0.93 | 12.32 | 0.96 |
| | | Type 2 | -8.13 | 0.93 | 10.89 | 0.96 |

Table 4.4 analyses the performance of the bootstrap estimators for estimating the variance of population-based R-indicators under the two extreme response rate scenarios, RR1 and RR3. Analytical expressions for the variance of sample-based R-indicators have been developed and used in the evaluation study (see Shlomo et al. 2012). Simulation means of the variance estimators are compared in Table 4.4 with the simulation variances (calculated across the replicated samples), using percentage

relative bias. The table also includes the Coverage Rate defined as the percentage of times that the true $R_\rho$ is contained in the confidence interval

$$100\left\{\left[\sum_{j=1}^{500} I\left(R_\rho \in \hat{R}_{\hat{\rho}j} \pm 1.96\sqrt{\hat{V}_j\left(\hat{R}_{\hat{\rho}j}\right)}\right)\right]/500\right\},$$

where $\hat{V}_j(\hat{R}_{\hat{\rho}j})$ is the estimated variance for the j-th sample (linearization variance estimator for sample-based estimator and bootstrap variance estimator for population-based estimators) and I is the indicator function. The bootstrap variance estimators for population-based estimators work well. The sample-based estimator show better coverage than the corresponding population-based versions. Type 1 and Type 2 population-based estimators have similar coverages. The coverage always improves as the sample size gets larger, except for the type 2 estimator under response rate RR1.

The behaviour under different response rates is mixed. There seems to be an interaction between sample size and response rate. The number of variables in the model does not have a large impact on coverage. Some difficulties are observed for population-based estimators, under the highest response rate (RR3) especially for 1% sample rate.

# 5. Application to the Dutch Health Survey

In this section, we apply the population-based Type 1 and Type 2 estimators to the Dutch Health Survey. We employ three auxiliary variables that are part of the gold standard for Dutch market research companies. The Health Survey is conducted by Statistics Netherlands, so that we can compare population-based performance to sample-based performance.

The Dutch Health Survey (HS) is commissioned since 1998 as a repeated cross-sectional survey among the full population registered in the Dutch Population Register, but excluding the institutionalized population. It uses a two-stage, self-weighting sampling design in which the first stage is formed by municipalities and the second stage by persons living in the selected municipalities. Until 2012, the HS was a face-to-face survey. In 2012, it changed to a mixed-mode design involving online and face-to-face interviews. Over the years, the sample size was reduced considerably from around 35,000 to around 18,000. We use the 2002 HS data, one of the last years with the original sample size. The net sample size is 33,584 persons. The response rate to the 2002 HS was 54.2%.

Table 5.1: Age, gender, and marital status distributions for the sample, respondents, and population.

| Variables | Categories | Respondents | Sample | Population |
|---|---|---|---|---|
| Age | 20-24 | 7.5 | 7.9 | 8.1 |
| | 25-29 | 7.3 | 8.2 | 8.9 |
| | 30-34 | 9.9 | 10.2 | 10.9 |
| | 35-39 | 10.9 | 10.8 | 11 |
| | 40-44 | 10.3 | 10.3 | 10.4 |
| | 45-49 | 9.7 | 9.4 | 9.6 |
| | 50-54 | 9.4 | 9.6 | 9.5 |
| | 55-59 | 8.8 | 8.9 | 8 |
| | 60-64 | 7.1 | 6.7 | 6.3 |
| | 65-69 | 5.9 | 5.6 | 5.4 |
| | 70-74 | 5.4 | 4.7 | 4.6 |
| | 75+ | 7.7 | 7.8 | 7.2 |
| Gender | Male | 48.9 | 49.8 | 49.2 |
| | Female | 51.1 | 50.2 | 50.8 |
| Marital status | Not married | 23.7 | 26.8 | 26.9 |
| | Married | 63.3 | 59.3 | 58.8 |
| | Widowed | 6.5 | 6.7 | 6.7 |
| | Divorced | 6.4 | 7.2 | 7.6 |

To calibrate national and regional samples, Dutch market research companies use the so-called Gold Standard population statistics produced by Statistics Netherlands (MOA 2015). The Gold Standard is an explicitly defined set of auxiliary variables that affiliated companies include in their survey questionnaires. Three of these variables are age, gender and marital status. We focus on these three in the application.

Table 5.1 contains the HS sample and response distributions, and the Statistics Netherlands' population distributions for the three variables. Joint population distributions, needed to estimate the Type 1 population-based covariance matrices, are also available, but not given here. In practice, the sample distribution is, of course, unknown. The three variables show a different picture: for age and marital status, the response distribution is closer to the sample distribution than to the population distribution, and population-based response propensities give more variation. For gender, the population distribution is closer to the response distribution and less variation is found.

We estimated Type 1 and Type 2 population-based R-indicators with and without the smoothing of the composite estimator. Table 5.2 contains the estimated smoothing parameter $\tilde{\lambda}_{opt}$ based on the population-based response propensities. We also include an estimate for $\lambda_{\mathrm{opt}}$ calculated using sample-based quantities. The latter can normally not be computed and is included for comparison only. The sample-based

$\tilde{\lambda}_{opt}$ are larger and tend to have a stronger smoothing effect. However, all $\tilde{\lambda}_{opt}$ are relatively small.

Table 5.3 contains the various population-based R-indicators. For comparison, the sample-based R-indicator is also provided where we used the logistic link function. The linear link function produced the same result. We can conclude that the population-based R-indicators, using only response and population distributions, are different from the sample-based R-indicators, using response and sample distributions. This difference increases, as expected, when Type 2 indicators are used. The composite estimators perform slightly better than the non-composite estimators, but there is still a considerable difference. This is not due to a biased smoothing parameter, as the difference is only modestly smaller when sample-based propensities are used to estimate the smoothing parameter. Furthermore, after bias adjustment, the differences between the composite estimators for sample-based and population-based propensities vanish.

Table 5.2: Values for lambda based on population-based response propensities and on sample-based response propensities for Type 1 and 2 composite estimators.

| | Smoothing parameter $\tilde{\lambda}_{opt}$ | |
|---|---|---|
| | Type 1 | Type 2 |
| Population-based response propensities | 0.043 | 0.038 |
| Sample-based response propensities | 0.076 | 0.095 |

Table 5.3: Unadjusted and bias-adjusted sample-based and Type 1 and Type 2 population-based R-indicators for the HS 2002 data. The population-based R – indicators are given without (original) and with composite estimator using population-based and sample-based response propensities. 95% confidence intervals (CI) by normal approximation are provided.

| | Unadjusted | | | Bias-adjusted | | |
|---|---|---|---|---|---|---|
| Estimator | R-ind | 95% CI | | R-ind | 95% CI | |
| Sample-based | 0.899 | 0.888 | 0.909 | 0.901 | 0.890 | 0.912 |
| Type 1 – original | 0.876 | 0.860 | 0.891 | 0.879 | 0.864 | 0.895 |
| Type 1 – composite population-based | 0.880 | 0.865 | 0.896 | 0.880 | 0.864 | 0.895 |
| Type 1 – composite sample-based | 0.883 | 0.868 | 0.898 | 0.880 | 0.865 | 0.895 |
| Type 2 - original | 0.873 | 0.858 | 0.889 | 0.877 | 0.861 | 0.894 |
| Type 2 – composite population-based | 0.878 | 0.863 | 0.894 | 0.878 | 0.862 | 0.893 |
| Type 2 – composite sample-based | 0.881 | 0.866 | 0.897 | 0.878 | 0.863 | 0.893 |

A conclusion from the application is that the lower population-based R-indicators result from the large differences between sample and population distributions of the auxiliary variables. For a sample size of 33,584 persons, the differences between sample and population distributions test as significant for all three variables at the 5% level. The available Dutch Health Survey net sample does not contain sampling units with frame and/or other administrative errors as well as out-of-scope populations such as institutionalized persons. This modification plus some additional, small tailoring to interviewer workloads, most likely caused sample distributions to differ from the original population counts. This points at the Achilles heel of population-based R-indicators; it is imperative that there is no disparity between definitions and populations.

# 6. Discussion

The extension of sample-based to population-based estimators of R-indicators is comprised of two steps: 1) the estimation of response propensities, and 2) the estimation of the R-indicators based on these propensities. The population-based estimation of response propensities is straightforward when linear models are assumed for response propensities and response influences. The linear link function is reasonable when estimating response propensities under typical response rates seen for large-scale national social surveys as shown in the evaluation study in Section 4. The sample-based estimators contain sample covariance matrices and sample frequencies that can be replaced by population covariance matrices or population frequencies. We identified two types: population cross-products are available or auxiliary information is restricted to marginal population counts only. We labelled the corresponding estimators as Type 1 and Type 2 estimators, respectively. The Type 2 setting is more restrictive than the Type 1 setting.

Following the estimation of population-based response propensities, we have constructed population-based estimators for the R-indicator and examined their properties both theoretically and empirically. The estimators are applied to samples drawn from real data from the 1995 Israel Census Data where ‘true’ propensities were calculated according to realistic assumptions of national household social surveys. Thus, we have addressed the first two research questions at the beginning of the paper: How to extend sample-based response propensities and  R-indicators to population-based response propensities and R-indicators? and What are the statistical properties of population-based R-indicators?.

There are many options for the estimation of R-indicators based on the response to the survey. We used propensity weighted response means as the propensities are available. However, any calibration method can be used such as linear weighting or adjustment classes. In fact, the set of auxiliary variables used for the estimation of the R-indicators may be a subset of the auxiliary variables used for the estimation of propensities and influences. Parsimonious models may prove to be more efficient as it is known that propensity-weighting may seriously affect the precision of the estimators. This is a topic for future research.

The two properties we examined are the bias and standard errors of the proposed population-based R-indicators. As expected the bias and standard errors are dependent on the size of the sample and the type of auxiliary information available where the smaller the sample, the larger the bias and the standard error. When samples are smaller, it becomes more difficult to distinguish sampling variation from response variation. Clearly, the confidence intervals become larger as there is less information in small samples.

The bias-adjusted Type 1 estimators (population cross-products) perform better than the bias-adjusted Type 2 estimators (population marginal counts). This is as expected given that they employ more information. However, the unadjusted Type 2 estimators have better RRMSE properties than the unadjusted Type 1 estimators. This is a surprising result and points at a suboptimal use of the population cross-products when they are used as 'plug-ins' and do not account for any sampling variation. The standard errors of the population-based estimators are larger than their sample-based counterparts.

For very high response rates as shown in the evaluation study (scenario RR3), the population-based R-indicators provide higher standard errors and larger bias, mainly due to propensities being estimated outside of the interval [0,1]. For this reason, we proposed a composite estimator with varying weights dependent on the response rate. Standard errors were reduced but at the cost of increased bias.

From the analyses it becomes apparent that the bias of the Type 1 and Type 2 estimators depends on the number of auxiliary variables, but this dependence was modest in our evaluations. When detailed models are used, containing many variables in the estimation of response propensities, then the bias may increase. The rationale behind this is that detailed models allow for more sampling variation to be picked up as bias.

The population-based R-indicators will have a number of caveats:

Firstly, we make the assumption that the survey measures the same quantities as in the population information and does not investigate the effect of possible departures from this assumption. However, we note that there is an imminent risk of measurement errors when comparing the representativeness of survey questions to population statistics. It must be ascertained that the survey questions that are employed have the same definitions and classifications as the population tables. Hence, it is best to avoid questions that are prone to measurement errors, such as questions that require a strong cognitive effort or that may lead to socially desirable answers. Also, it is strongly recommended to use population statistics that are based on registrations or administrative data. The population-based R-indicators can be used for population statistics that are based on surveys, but these statistics may not reflect the true population distribution accurately. One would draw erroneous conclusions about the representativeness of the response if the population estimates are biased.

Secondly, in settings where only population information is available, options to improve representativeness during data collection are much more limited; for the nonrespondents no individual auxiliary information is available. Nonetheless, in these settings, assessments of representativeness may still be useful in the design of advance and reminder letters, in interviewer training and in paradata collection.

Finally, in this paper, we do not consider hybrid settings where the R-indicator is based on both linked data and population tables. Such an extension is relatively straightforward but will be left to future papers.

The research into population-based indicators is still at the beginning stage and it is too early to provide a definitive answer to the last research question presented in the introduction regarding the feasibility and practicability of R-indicators based on aggregate population auxiliary information. The evaluation study presented in Section 4 is based on real data under realistic assumptions of response probabilities typically found in social surveys conducted at statistical agencies. Future research needs to assess whether alternative estimators can be constructed that are more precise, and, consequently, allow for stronger conclusions regarding the nature of response. A natural avenue to explore is a modification of the EM-algorithm, in which the score of the nonrespondents on the auxiliary variables are estimated and used to update response propensity estimates.

We did not consider population-based estimation for other types of models such as logistic or probit regression. As shown in the numerical evaluation in Section 4, differences in sample-based estimators between the linear and logistic link function are in general small, but when the response rates get very close to 1, they become more evident. For these cases, developing other link functions for population-based estimation is a subject of future research. This would be a useful and natural extension to the theory of R-indicators as these models are often used in practice and avoid propensities outside the [0, 1] interval. Also for these extensions, it would be advisable to move to an iterative approach such as the EM-algorithm.

Although the R-indicators were motivated in this paper from survey data collection practice, they can be applied to any setting with missing data on variables of interest and (almost) complete auxiliary data. They can, for instance, be used to monitor and evaluate the completion of administrative data, which is useful when administrative data have a time lag and fill gradually over time.

## Acknowledgements

Part of the research presented here was developed within project RISQ (Representativity Indicators for Survey Quality, www.risq-project.eu), funded by the European 7th Framework Programme. We thank the members of the RISQ project: Katja Rutar from Statistični Urad Republike Slovenije, Geert Loosveldt and Koen Beullens from Katholieke Universiteit, Leuven, Øyvin Kleven, Johan Fosen and Li-Chun Zhang from Statistisk Sentralbyrå, Norway, Ana Marujo from the University of Southampton, UK and Paul Knottnerus, Centraal Bureau voor de Statistiek, for their valuable input.
The first author was supported by a STSM Grant from the COST Action IS1004 and by the ex 60% University of Bergamo, Biffignandi grant.

# References

Bethlehem, J. (1988), Reduction of Nonresponse Bias Through Regression Estimation, Journal of Official Statistics, 4, 251-260.

Booth, J.G., Butler, R.W. and Hall, P. (1994), Bootstrap Methods for Finite Populations, Journal of the American Statistical Association, 89 (428), 1282-1289.

Copas, J.B. (1983), Regression, prediction and shrinkage, Journal of the Royal Statistical Society, Series B, 45, 311 – 354.

Copas, J.B. (1993), The shrinkage of point scoring methods,  Journal of the Royal Statistical Society, Series C, 42, 315 – 331.

De Heij, V., Schouten, B. and Shlomo, N. (2015), RISQ manual 2.1. Tools in SAS and R for the computation of R-indicators and partial R-indicators, available at www.risq-project.eu.

Deville, J.C. and Sarndal, C.E. (1992), Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376-382.

Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap. New York: Chapman and Hall.

Kreuter, F. (2013), Improving surveys with process and paradata, Edited monograph, John Wilen and Sons, Hoboken, New Jersey, USA.

Little, R.J.A. (1986), Survey nonresponse adjustments for estimates of means. International Statistical Review, 54, 139-157.

Little, R.J.A. (1988), Missing-data adjustments in large surveys. Journal of Business and Economic Statistics, 6, 287-301.

Little, R.J.A. and Rubin, D.B. (2002), Statistical Analysis with Missing Data, Hoboken, New Jersey: Wiley.

Lundquist, P., Särndal, C.E. (2013), Aspects of responsive design with applications to the Swedish Living Conditions Survey, Journal of Official Statistics, 29 (4), 557 – 582.
MOA (2015), User Instruction Gold Standard, Dutch Market Research Association, available at www.moaweb.nl/sevrices/services/gouden-standaard.html.

Rosenbaum, P.R. and Rubin, D.B. (1983), The central role of the propensity score in observational studies for causal effects, Biometrika, 70, 41-55.

Särndal, C.E. (2011), The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation, Journal of Official Statistics, 27 (1), 1 – 21.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling, Springer, New York.

Särndal, C.E. and P. Lundquist (2014), Accuracy in Estimation with Nonresponse: A Function of Degree of Imbalance and Degree of Explanation, Journal of Survey Statistics and Methodology, 2 (4), 361 – 387.

Särndal, C.E. and Lundström, S. (2005), Estimation in Surveys with Nonresponse, New York, Wiley.

Schouten, B., Bethlehem, J., Beulens, K., Kleven, Ø., Loosveldt, G., Rutar, K., Shlomo, N., and Skinner, C. (2012), Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators, International Statistical Review, 80 (3), 382 – 399.

Schouten, B., Calinescu, M. and Luiten, A. (2013), Optimizing quality of response through adaptive survey designs, Survey Methodology, 39 (1), 29 – 58.

Schouten, B., Cobben, F. and Bethlehem, J. (2009), Indicators for the Representativeness of Survey Response. Survey Methodology, 35, 101-113.

Schouten, B., Shlomo, N., Skinner, C. (2011), Indicators for Monitoring and Improving Representativeness of Response. J. Off. Stat. 27, 231—253.

Schouten, B., Cobben, F., Lundquist, P. and Wagner, J. (2016), Does more balanced survey response imply less non-response bias?, Journal of the Royal Statistical Society, Series A, 179 (3), 727-748.

Shlomo, N., Skinner, C., and Schouten, B. (2012), Estimation of an Indicator of the Representativeness of Survey Response. Journal of Statistical Planning and Inference, 142, 201-211.

Wagner, J. (2012), A comparison of alternative indicators for the risk of nonresponse bias, Public Opinion Quarterly, 76 (3), 555 – 575.

Wagner, J. (2013), Adaptive contact strategies in telephone and face-to-face surveys, Survey Research Methods, 7 (1), 45 – 55.

Wagner, J. and Hubbard, F. (2014), Producing unbiased estimates of propensity models during data collection, Journal of Survey Statistics and Methodology, 2, 323 – 342.

Wolter, K.M. (2007), Introduction to Variance Estimation. 2nd Ed. New York: Springer.

# Appendix A - Analytic approximation to the bias of Type 1 estimators

First, we compute the bias of $\tilde{S}^2_{\tilde{\rho}_{T1}}$ under general sampling design. Letting

$$\hat{m}_1 = N^{-1} \sum\nolimits_r d_i$$
$$\hat{m}_2 = N^{-1} \sum\nolimits_r d_i \tilde{\rho}_{i,T1}$$

we can write

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = E\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) - S^2_{\rho} = \frac{N}{N-1}\left\{E\left(\hat{m}_2\right) - V\left(\hat{m}_1\right) - \left[E\left(\hat{m}_1\right)\right]^2\right\} - \frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in U}\rho_i^2 - \bar{\rho}_U^2\right\}. \tag{A1}$$

Note that

$$E\left(\hat{m}_2\right) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i \tilde{\rho}_{i,T1}\right) = \frac{1}{N}\sum_{i\in U} \mathbf{x}_i^T \mathbf{T}_1^{-1} E_s\left\{E_m\left(d_i^2 s_i r_i \mathbf{x}_i + \sum_{\substack{k\in U\\ k\neq i}} d_i d_k s_i s_k r_i r_k \mathbf{x}_k \middle| s\right)\right\}$$
$$= \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1} \mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1} \sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik} \rho_k \mathbf{x}_k,$$
$$E\left(\hat{m}_1\right) = E\left(\frac{1}{N}\sum_{i\in U} d_i s_i r_i\right) = E_s\left(\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right) = \bar{\rho}_U,$$

and

$$V\left(\hat{m}_1\right) = V_s\left\{E_m\left(\hat{m}_1 \middle| s\right)\right\} + E_s\left\{V_m\left(\hat{m}_1 \middle| s\right)\right\}$$
$$= V_s\left\{\frac{1}{N}\sum_{i\in U} d_i s_i \rho_i\right\} + E_s\left\{\frac{1}{N^2}\sum_{i\in U} d_i^2 s_i \rho_i\left(1-\rho_i\right)\right\}$$
$$= \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik} \rho_i \rho_k + \frac{1}{N^2}\sum_{i\in U} d_i \rho_i\left(1-\rho_i\right),$$

where $\Delta_{ik} = \pi_{ik} - \pi_i \pi_k$ and $\pi_{ik}$ are the second-order sample inclusion probabilities. Hence, the bias of $\tilde{S}^2_{\tilde{\rho}_{T1}}$ with respect to the joint distribution of sampling design and the response mechanism is given by

$$B\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right) = \frac{N}{N-1}\left[\frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1} \mathbf{x}_i + \frac{1}{N}\sum_{i\in U} d_i \rho_i \mathbf{x}_i^T \mathbf{T}_1^{-1} \sum_{\substack{k\in U\\ k\neq i}} d_k \pi_{ik} \rho_k \mathbf{x}_k - \frac{1}{N^2}\sum_{i\in U}\sum_{k\in U} d_i d_k \Delta_{ik} \rho_i \rho_k \right.$$
$$\left. - \frac{1}{N^2}\sum_{i\in U} d_i \rho_i\left(1-\rho_i\right) - \frac{1}{N}\sum_{i\in U}\rho_i^2\right]. \tag{A2}$$

Under simple random sampling without replacement, (A2) can be simplified to

$$B^{SRS}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)=\frac{N}{N-1}\left[\frac{1}{n}\sum_{i\in U}\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\}\boldsymbol{x}_i^T\boldsymbol{T}_1^{-1}\boldsymbol{x}_i+\frac{n-1}{n(N-1)N}\sum_{i\in U}\rho_i^2-\frac{\bar{\rho}_U}{n}-\left(1-\frac{n}{N}\right)\frac{S^2_\rho}{n}\right].$$

A response-set based estimator of $B^{SRS}(\tilde{S}^2_{\tilde{\rho}T_1})$ is

$$\tilde{B}^{SRS}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)=\frac{N}{N-1}\left[\frac{N}{n^2}\sum_{i\in r}\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T1}\right\}\boldsymbol{x}_i^T\boldsymbol{T}_1^{-1}\boldsymbol{x}_i+\frac{n-1}{n^2(N-1)}\sum_{i\in r}\tilde{\rho}_{i,T1}-\left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T1}}}{n}-\frac{n_r}{n^2}\right].$$

More generally, the Horwitz-Thompson response-set estimator for (A2) under complex sampling is given by

$$\tilde{B}_{\tilde{\rho}_{T1}}\left(\tilde{S}^2_{\tilde{\rho}_{T1}}\right)=\frac{N}{N-1}\left\{\frac{1}{N}\sum_{i\in r}d_i\left(d_i-\tilde{\rho}_{i,T1}\right)\boldsymbol{x}_i^T\boldsymbol{T}_1^{-1}\boldsymbol{x}_i-\frac{1}{N^2}\sum_{i\in r}d_i^3\Delta_{ii}\tilde{\rho}_{i,T1}-\frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\k\neq i}}d_id_k\frac{\Delta_{ik}}{\pi_{ik}}\right.$$
$$\left.-\frac{1}{N^2}\sum_{i\in r}d_i^2\left(1-\tilde{\rho}_{i,T1}\right)+\frac{1}{N}\sum_{i\in r}x_i^TT_1^{-1}\sum_{\substack{k\in r\\k\neq i}}x_k\left(d_id_k-\frac{1}{\pi_{ik}}\right)\right\}.$$

# Appendix B - Analytic approximation to the bias of Type 2 estimators

The strategy to compute an analytical bias adjustment for $\tilde{S}^2_{\tilde{\rho}T_2}$ is to first approximate $\tilde{\rho}_{i,T2}$ by a linear estimator using Taylor linearization techniques. Next, compute an approximate bias adjustment for $\tilde{S}^2_{\tilde{\rho}T_2}$, by inserting the linear approximation for $\tilde{\rho}_{i,T2}$ into $\widehat{m}_2$.

In the following, define, for j=1,…p and j'=1,…p, the estimated totals

$$\hat{t}_0=\sum\nolimits_s d_kr_k,\qquad \hat{t}_{jj'}=\sum\nolimits_s d_kr_kz_{jk}z_{j'k},\qquad \text{and}\qquad \hat{t}_j=\sum\nolimits_s d_kr_kx_{jk},$$

where

$$\boldsymbol{z}_k=\left(\boldsymbol{x}_k-\bar{\boldsymbol{x}}_U\right)$$
$$z_{jk}=\left(x_{jk}-\bar{x}_{jU}\right).$$

Let $\hat{t}$ be a p-vector with components $\hat{t}_j$, and $\hat{F}$ be the symmetric $(p \times p)$-matrix with elements $\hat{t}_{jj'}$. We may write

$$\tilde{\rho}_{i,T2} = \boldsymbol{x}_i^T \left[ N\hat{t}_0^{-1}\hat{\boldsymbol{F}} + N\overline{\boldsymbol{x}}_U \overline{\boldsymbol{x}}_U^T \right]^{-1} \hat{\boldsymbol{t}} = \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{t}} .$$

Now, define the population totals

$$t_0 = \sum_U \rho_k , \quad \boldsymbol{F} = \sum_U \rho_k \boldsymbol{z}_k \boldsymbol{z}_k^T , \quad \text{and} \quad \boldsymbol{t} = \sum_U \rho_k \boldsymbol{x}_k .$$

Notice that $\hat{t}_0$ is unbiased for $t_0$, $\hat{F}$ is unbiased for $F$, and $\hat{t}$ is unbiased for $t$. Let

$$\boldsymbol{T}_2 = Nt_0^{-1}\boldsymbol{F} + N\overline{\boldsymbol{x}}_U \overline{\boldsymbol{x}}_U^T .$$

*Proposition 1*: *The estimator* $\tilde{\rho}_{i,T2}$ *defined in (7) may be approximated by*

$$\tilde{\rho}_{i,T2} \cong \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \left( Nt_0^{-2}\boldsymbol{F} \right) \boldsymbol{T}_2^{-1} \boldsymbol{t} \left( \hat{t}_0 - t_0 \right) - \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} Nt_0^{-1} \left( \hat{\boldsymbol{F}} - \boldsymbol{F} \right) \boldsymbol{T}_2^{-1} \boldsymbol{t} + \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \hat{\boldsymbol{t}} .$$

Proof. Following standard Taylor linearization (see Särndal, Swensson and Wretman 1992 and Bethlehem 1988), the estimator $\tilde{\rho}_{i,T2}$ may be approximated by

$$\tilde{\rho}_{i,T2} \cong \rho_{i,T2}^{*} + a_0 \left( \hat{t}_0 - t_0 \right) + \sum_{j=1}^{p} \sum_{j' \le j} a_{jj'} \left( \hat{t}_{jj'} - t_{jj'} \right) + \sum_{j=1}^{p} a_j \left( \hat{t}_j - t_j \right), \tag{A3}$$

where

$$\rho_{i,T2}^{*} = \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \boldsymbol{t} ,$$

and

$$a_0 = \left. \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_0} \right|_{\substack{\hat{t}_0 = t_0 \\ \hat{\boldsymbol{F}} = \boldsymbol{F} \\ \hat{\boldsymbol{t}} = \boldsymbol{t}}} = \boldsymbol{x}_i^T \left[ -\hat{\boldsymbol{T}}_2^{-1} \left( -N\hat{t}_0^{-2}\hat{\boldsymbol{F}} \right) \hat{\boldsymbol{T}}_2^{-1} \right] \hat{\boldsymbol{t}} \Big|_{\substack{\hat{t}_0 = t_0 \\ \hat{\boldsymbol{F}} = \boldsymbol{F} \\ \hat{\boldsymbol{t}} = \boldsymbol{t}}} = \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \left( Nt_0^{-2}\boldsymbol{F} \right) \boldsymbol{T}_2^{-1} \boldsymbol{t},$$

$$a_{jj'} = \left. \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_{jj'}} \right|_{\substack{\hat{t}_0 = t_0 \\ \hat{\boldsymbol{F}} = \boldsymbol{F} \\ \hat{\boldsymbol{t}} = \boldsymbol{t}}} = -\boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \left( Nt_0^{-1}\boldsymbol{\Lambda}_{jj'} \right) \boldsymbol{T}_2^{-1} \boldsymbol{t},$$

$$a_j = \left. \frac{\partial \tilde{\rho}_{i,T2}}{\partial \hat{t}_j} \right|_{\substack{\hat{t}_0 = t_0 \\ \hat{\boldsymbol{F}} = \boldsymbol{F} \\ \hat{\boldsymbol{t}} = \boldsymbol{t}}} = \boldsymbol{x}_i^T \boldsymbol{T}_2^{-1} \boldsymbol{\lambda}_j ,$$

where $\Lambda_{jj'}$ is a $(p \times p)$-matrix with ones in positions $(j, j')$ and $(j, j')$ and zeros elsewhere and $\lambda_j$ is a p-vector with the j-th component equal to one and zeros elsewhere. Inserting the partial derivatives into (A3) gives the result.

□

*Proposition 2. Under simple random sampling, an approximate bias for* $\tilde{S}^2_{\tilde{\rho}_{T2}}$ *with respect to the joint distribution of sampling design and the response mechanism is given by*

$$B^{SRS}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\left\{ t_0^{-2}\frac{N}{n}\sum_U c_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} - t_0^{-1}\frac{N}{n}\sum_U \boldsymbol{b}_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\}\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{t}\right.$$
$$+\frac{1}{n}\sum_U \rho_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{x}_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} + \frac{n-1}{n(N-1)}\sum_U \rho_i\rho^*_{i,T2} - \left(1-\frac{n}{N}\right)\frac{S^2_\rho}{n} - \frac{\overline{\rho}_U}{n}$$
$$\left.+\frac{1}{nN}\sum_U \rho_i^2 - \frac{1}{N}\sum_U \rho_i^2\right\},$$

where

$$c_i = \boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{F}\boldsymbol{T}_2^{-1}\boldsymbol{t},$$
$$\boldsymbol{b}_i = \boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}$$
$$\rho^*_{i,T2} = \boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{t}.$$

A response-set based estimator of $B^{SRS}(\tilde{S}^2_{\tilde{\rho}T2})$ is

$$\tilde{B}^{SRS}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1}\left\{\frac{1}{n_r^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{F}}\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{t}} - \frac{N}{nn_r}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\boldsymbol{z}_i\boldsymbol{z}_i^T\hat{\boldsymbol{T}}_2^{-1}\hat{\boldsymbol{t}}\right.$$
$$\left.+\frac{N}{n^2}\sum_r\left\{1-\frac{n-1}{N-1}\tilde{\rho}_{i,T2}\right\}\boldsymbol{x}_i^T\hat{\boldsymbol{T}}_2^{-1}\boldsymbol{x}_i + \frac{n-1}{n^2(N-1)}\sum_r\tilde{\rho}_{i,T2} - \left(1-\frac{n}{N}\right)\frac{\tilde{S}^2_{\tilde{\rho}_{T2}}}{n} - \frac{n_r}{n^2}\right\}.$$

Proof. Thanks to Proposition 1, $\hat{m}_2$ defined in Appendix A may be approximated as follows

$$\begin{aligned}\hat{m}_2 &= \frac{1}{N}\sum_U d_is_ir_i\tilde{\rho}_{i,T2}\\ &\cong \frac{1}{N}\sum_U d_is_ir_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\left(Nt_0^{-2}\boldsymbol{F}\right)\boldsymbol{T}_2^{-1}\boldsymbol{t}\left(\hat{t}_0-t_0\right) - \frac{1}{N}\sum_U d_is_ir_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}Nt_0^{-1}\left(\hat{\boldsymbol{F}}-\boldsymbol{F}\right)\boldsymbol{T}_2^{-1}\boldsymbol{t} + \frac{1}{N}\sum_U d_is_ir_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\hat{\boldsymbol{t}}\\ &=: A+B+C.\end{aligned}$$

The expected values of the terms A, B, and C are

$$E(A) = t_0^{-2}\sum_{i\in U}c_id_i\rho_i + t_0^{-2}\sum_{i\in U}c_id_i\sum_{k\neq i}d_k\rho_i\rho_k\pi_{ik} - t_0^{-1}\sum_{i\in U}c_i\rho_i,$$
$$E(B) = -t_0^{-1}\sum_{i\in U}d_ib_i\rho_i\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{t} - t_0^{-1}\sum_{i\in U}d_i\boldsymbol{b}_i\sum_{k\neq i}d_k\rho_i\rho_k\pi_{ik}\boldsymbol{z}_k\boldsymbol{z}_k^T\boldsymbol{T}_2^{-1}\boldsymbol{t} + t_0^{-1}\sum_{i\in U}\rho_i\boldsymbol{b}_i\boldsymbol{F}\boldsymbol{T}_2^{-1}\boldsymbol{t},$$

and

$$E(C) = \frac{1}{N}\sum_{i\in U}d_i\rho_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{x}_i + \frac{1}{N}\sum_{i\in U}d_i\rho_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\sum_{k\neq i}d_k\rho_k\pi_{ik}\boldsymbol{x}_k.$$

It follows that, under simple random sampling, $E(\hat{m}_2)$ becomes

$$E^{SRS}(\hat{m}_2) = t_0^{-2}\frac{N}{n}\sum_U c_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} - t_0^{-1}\frac{N}{n}\sum_U\boldsymbol{b}_i\rho_i\left\{1-\frac{n-1}{N-1}\rho_i\right\}\boldsymbol{z}_i\boldsymbol{z}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{t}$$
$$+\frac{1}{n}\sum_U\rho_i\boldsymbol{x}_i^T\boldsymbol{T}_2^{-1}\boldsymbol{x}_i\left\{1-\frac{n-1}{N-1}\rho_i\right\} + \frac{n-1}{n(N-1)}\sum_U\rho_i\rho^*_{i,T2}.$$

So the total bias under simple random sampling is obtained by inserting $E^{SRS}(\widehat{m}_2)$ computed above into (A1) and following the proof in Appendix A for the other terms. The response-set based estimator $\tilde{B}^{SRS}_{\tilde{\rho}_{T2}}(\tilde{S}^2_{\tilde{\rho}_{T2}})$ of $B^{SRS}(\tilde{S}^2_{\tilde{\rho}_{T2}})$ is obtained by substituting $t_0$ with $\hat{t}_0 = Nn_r/n$, $F$ with

$$\hat{\boldsymbol{F}} = Nn^{-1} \sum\nolimits_r \boldsymbol{z}_k \boldsymbol{z}_k^T ,$$

$T_2$ with

$$\hat{\boldsymbol{T}}_2 = N\hat{t}_0^{-1} \hat{\boldsymbol{F}} + N\overline{\boldsymbol{x}}_U \overline{\boldsymbol{x}}_U^T ,$$

and $t$ with

$$\hat{\boldsymbol{t}} = Nn^{-1} \sum\nolimits_r \boldsymbol{x}_k$$

□

Note that the bias adjustment $\tilde{B}^{SRS}_{\tilde{\rho}_{T2}}(\tilde{S}^2_{\tilde{\rho}_{T2}})$ corresponds to 'plugging-in' Type 2 quantities ($\tilde{\rho}_{i,T2}$ instead of $\tilde{\rho}_{i,T1}$, matrix $\hat{T}_2$ instead of $T_1$, and $\tilde{S}^2_{\tilde{\rho}_{T2}}$ instead of$\tilde{S}^2_{\tilde{\rho}_{T1}}$) into the analytical bias adjustment $\tilde{B}^{SRS}_{\tilde{\rho}_{T1}}(\tilde{S}^2_{\tilde{\rho}_{T1}})$ developed for $\tilde{S}^2_{\tilde{\rho}_{T1}}$ with two additional terms due to the linearization of $\hat{T}_2$.

More generally, the Horwitz-Thompson response-set estimator under complex sampling for the bias adjustment of Type 2 population-based R-indicator is given by

$$\begin{aligned}
\tilde{B}_{\tilde{\rho}_{T2}}\left(\tilde{S}^2_{\tilde{\rho}_{T2}}\right) = \frac{N}{N-1} \Bigg\{ & \frac{1}{N}\sum_{i\in r} d_i \left(d_i - \tilde{\rho}_{i,T2}\right) \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \boldsymbol{x}_i - \frac{1}{N^2}\sum_{i\in r} d_i^3 \Delta_{ii} \tilde{\rho}_{i,T2} - \frac{1}{N^2}\sum_{i\in r}\sum_{\substack{k\in r\\ k\neq i}} d_i d_k \frac{\Delta_{ik}}{\pi_{ik}} \\
& - \frac{1}{N^2}\sum_{i\in r} d_i^2 \left(1 - \tilde{\rho}_{i,T2}\right) + \frac{1}{N}\sum_{i\in r} x_i^T \hat{T}_2^{-1} \sum_{\substack{k\in r\\ k\neq i}} x_k \left( d_i d_k - \frac{1}{\pi_{ik}} \right) + \left(\sum_{k\in r} d_k\right)^{-2} \sum_{i\in r} d_i^2 \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{F}} \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{t}} \\
& + \left(\sum_{k\in r} d_k\right)^{-2} \sum_{i\in r} d_i \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{F}} \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{t}} \sum_{k\neq i} d_k - \left(\sum_{k\in r} d_k\right)^{-1} \sum_{i\in r} d_i^2 \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \boldsymbol{z}_i \boldsymbol{z}_i^T \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{t}} \\
& - \left(\sum_{k\in r} d_k\right)^{-1} \sum_{i\in r} d_i \boldsymbol{x}_i^T \hat{\boldsymbol{T}}_2^{-1} \sum_{k\neq i} d_k \boldsymbol{z}_k \boldsymbol{z}_k^T \hat{\boldsymbol{T}}_2^{-1} \hat{\boldsymbol{t}} \Bigg\}.
\end{aligned}$$

## Explanation of symbols

| | |
|---|---|
| Empty cell | Figure not applicable |
| . | Figure is unknown, insufficiently reliable or confidential |
| * | Provisional figure |
| ** | Revised provisional figure |
| 2014–2015 | 2014 to 2015 inclusive |
| 2014/2015 | Average for 2014 to 2015 inclusive |
| 2014/'15 | Crop year, financial year, school year, etc., beginning in 2014 and ending in 2015 |
| 2012/'13–2014/'15 | Crop year, financial year, etc., 2012/'13 to 2014/'15 inclusive |

Due to rounding, some totals may not correspond to the sum of the separate figures.

## Colofon

*Publisher*
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

*Prepress*
Statistics Netherlands, Studio BCO

*Design*
Edenspiekermann

*Information*
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2016.
Reproduction is permitted, provided Statistics Netherlands is quoted as the source.