Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2016

•

Volume 42

•

Number 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (statcan.smj-rte.statcan@canada.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 42, Number 2, December 2016

Contents

Regular papers

Short notes

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 195-218
Statistics Canada, Catalogue No. 12-001-X

Tests for evaluating nonresponse bias in surveys

Sharon L. Lohr, Minsun K. Riddles and David Morganstein[1]

Abstract

How do we tell whether weighting adjustments reduce nonresponse bias? If a variable is measured for everyone in the selected sample, then the design weights can be used to calculate an approximately unbiased estimate of the population mean or total for that variable. A second estimate of the population mean or total can be calculated using the survey respondents only, with weights that have been adjusted for nonresponse. If the two estimates disagree, then there is evidence that the weight adjustments may not have removed the nonresponse bias for that variable. In this paper we develop the theoretical properties of linearization and jackknife variance estimators for evaluating the bias of an estimated population mean or total by comparing estimates calculated from overlapping subsets of the same data with different sets of weights, when poststratification or inverse propensity weighting is used for the nonresponse adjustments to the weights. We provide sufficient conditions on the population, sample, and response mechanism for the variance estimators to be consistent, and demonstrate their small-sample properties through a simulation study.

Key Words: Inverse propensity weighting; Poststratification; Replication variance estimation; Responsive design.

1 Introduction

Nonresponse rates in probability samples are increasing worldwide. The U.S. Office of Management and Budget requires a nonresponse bias analysis when response rates are low or there are other indications that bias may be a problem (United States Office of Management and Budget 2006). Groves (2006) recommended using multiple approaches to assess potential nonresponse bias on key survey estimates.

Assessing potential nonresponse bias typically requires an external "gold standard" data source or rich sampling frame information. Common approaches for assessing nonresponse bias include: (1) comparing frame variables for respondents and nonrespondents, (2) comparing early and late respondents on frame variables and key survey variables, and (3) comparing estimates from the survey respondents (using nonresponse-adjusted weights) with estimates from an independent gold standard source. Differences in (1) and (2), however, do not necessarily imply that nonresponse bias remains after the weights are adjusted through calibration or propensity methods. If weight adjustments such as those described in Brick (2013) are successful in adjusting for nonresponse bias, the estimates from the survey using the nonresponse-adjusted weights may be approximately unbiased even if assessments (1) and (2) show differences.

In this paper we compare an estimate calculated using base weights from the selected sample with an estimate of the same quantity calculated using nonresponse-adjusted weights from the respondents only. An example might be comparing the estimated proportion of persons living in census tracts with more than 50% of housing units being owner occupied from (1) the selected sample, using the base weights, (2) the respondents, using the base weights, and (3) the respondents, using nonresponse-adjusted and/or poststratified weights. All three estimates of the proportion use the same characteristic, y, which is assumed to be known for everyone in the selected sample.

1. Sharon Lohr and David Morganstein are Vice Presidents and Minsun Riddles is Statistician at Westat, 1600 Research Blvd., Rockville MD 20850. E-mail: sharonlohr@westat.com.

The requirement that y be known for the selected sample restricts the set of variables that can be used to test for nonresponse bias. Typically, many of the key variables of interest are available only for the respondents, not for the entire selected sample. Other variables that are available for the entire selected sample may be used for poststratification or other nonresponse weighting adjustments. Poststratification forces the estimates of population totals for poststratification variables to equal the independent population counts for these variables, so these variables would not be expected to exhibit nonresponse bias after weight adjustments are performed. Variables that are available for the entire selected sample but are not used in the nonresponse weighting adjustments, and variables that are correlated with key survey variables, are the best choices for testing nonresponse bias. Examples of such variables include sample frame variables that are not used in poststratification (for example, an e-mail survey of university students may have information on academic performance that is not used in the nonresponse weighting), characteristics from a census (such as percent poverty in the block containing the sampled address), or information gathered by the interviewer (such as indications of children in the household that are visible from the street).

Eltinge (2002) and Harris-Kojetin (2012) recommended comparing estimates using different sets of weights to assess nonresponse bias and to choose among competing sets of nonresponse-adjusted weights. Such comparisons are common in nonresponse bias analyses: for example, Hamrick (2012) compared respondents with the full sample in the Eating and Health Module of the American Time Use Survey. To date, however, there has been no comprehensive examination of the statistical properties underlying these comparisons. In this paper, we derive the theoretical properties of variance estimators and hypothesis tests for the differences among estimated means that are calculated using the same outcome variable but with different weights and subsets of the data, and give conditions that will ensure consistency of the variance estimators.

Poststratification or inverse propensity weighting are commonly used to compensate for nonresponse bias. Yung and Rao (2000) derived linearization and jackknife estimators for the variance of a population mean estimated using poststratification, with and without nonresponse. They considered a uniform response mechanism in which each poststratum has the same response propensity, and considered the response indicator to be a fixed characteristic of the finite population. Kim and Kim (2007) studied asymptotic properties for inverse propensity weight adjustments, assuming that the response indicators of different units are independent. The previous work studied the variance of the poststratified or inverse-propensity-weighted statistic of interest. The problem we consider differs from the previous work because the estimated population total from the selected sample is often highly correlated with the estimate calculated using the respondents only, particularly when the response rate approaches one. The linearization and replication variance estimators in this paper account for that high correlation between the two sets of estimates, and thus can be used for testing the hypothesis that the poststratification or inverse propensity weighting removes the bias for the variables studied. We also extend previous research by allowing the response indicators to be correlated within primary sampling units, reflecting possible within-cluster homogeneity for responding to the survey.

Section 2 defines the parameter to be tested in the poststratification setting, derives the linearization and jackknife variance estimators, and gives sufficient conditions for the variance estimators to be consistent. In some circumstances the linearized variance of the test statistic may be zero under the null hypothesis, in

which case higher-order terms of the variance are needed. The higher-order terms are derived for the special case of simple random sampling in Theorem 3. Section 3 provides the linearization and jackknife variance estimators for testing the hypothesis that the propensity weights remove the nonresponse bias. Section 4 presents simulation studies and Section 5 contains concluding remarks and discusses future work.

2 Poststratification

2.1 Parameter and linearization variance

Suppose the finite population U has H strata, with N_h primary sampling units (PSUs) in stratum h, M_{hi} units in PSU i of stratum h, and $M = \sum_{hi} M_{hi}$ units in total. Let y_{hik} denote the quantity of interest for unit k in PSU (hi). A probability sample S is taken from the population, with n_h PSUs selected from stratum h and $n = \sum_{h=1}^{H} n_h$. The sample of PSUs from stratum h is denoted by S_h, and the sample of units from PSU (hi) is denoted by S_{hi}. Each unit has a design weight $w_{hik} = 1/P(\text{unit } hik \in S)$, and the PSU-level design weight is $w_{hi} = 1/P(\text{PSU } hi \in S_h)$.

Two frameworks are commonly used for the nonresponse mechanism. In a two-phase "forward" framework, the sample is selected at phase 1 and the nonresponse mechanism is a second phase of selection (Oh and Scheuren 1987; Särndal and Lundström 2005). Fay (1991) proposed a "reverse framework" which was studied further by Shao and Steel (1999) and Haziza, Thompson, and Yung (2010). In this framework, the nonresponse mechanism is applied to the finite population first, and then the sample is selected. The reverse framework, which we follow in this paper, specifies a nonresponse mechanism for nonsampled as well as sampled units. We assume that every unit in the population has a value of the response indicator r_{hik}. Let $R_{hik} = E[r_{hik}]$ under the response mechanism in the finite population, so that R_{hik} is the value of the true response propensity for unit (hik) in the population.

Suppose the characteristic y is known for all units in the selected sample. We compare the estimated population total using everyone in the sample with the estimated total using the poststratification-weighted respondents. There are C poststrata and poststratum c has M_c population units with $M = \sum_{c=1}^{C} M_c$. The poststratum counts M_c may be obtained from the sampling frame if the poststratification variables are known for every unit in the frame. Often, however, the poststratum counts come from an external source such as a census. Let $\delta_{chik} = 1$ if unit (hik) is in poststratum c and 0 otherwise. The population response rate in poststratum c is $p_c = \sum_{hik \in U} R_{hik}\delta_{chik} / M_c$. Yung and Rao (2000) assumed that the response rate p_c was the same for each poststratum. In many applications, however, the poststrata are formed so that response propensities within each poststratum are homogeneous, but the poststrata themselves have different mean response propensities. We therefore allow p_c to differ among the poststrata.

If y is known for all members of the selected sample, then the estimator of the population total using the sample is

$$\hat{Y}_{SS} = \sum_{hik \in S} w_{hik} y_{hik} = \sum_{hik \in U} Z_{hik} w_{hik} y_{hik}, \tag{2.1}$$

where w_{hik} is the design weight for unit k of PSU i in stratum h and Z_{hik} is the indicator variable for sample inclusion. Using the respondents only, the poststratified estimator of the population total is

$$\hat{Y}_{PS} = \sum_{c=1}^{C} M_c \frac{\sum_{hik \in S} w_{hik} r_{hik} \delta_{chik} y_{hik}}{\sum_{hik \in S} w_{hik} r_{hik} \delta_{chik}} = \sum_{c=1}^{C} M_c \frac{\hat{Y}_c^R}{\hat{M}_c^R}. \tag{2.2}$$

We define the finite population parameter of interest to be the difference between the expected value of $\hat{Y}_{PS}$ and the expected value of $\hat{Y}_{SS}$, which will be 0 if there is no nonresponse bias after poststratification. Define

$$\begin{aligned} M_c^R &= \sum_{hik \in U} \delta_{chik} R_{hik} = p_c M_c, \\ Y_c^R &= \sum_{hik \in U} \delta_{chik} R_{hik} y_{hik}, \end{aligned}$$

and

$$\theta = \sum_{c=1}^{C} M_c \frac{Y_c^R}{M_c^R} - Y = \sum_{c=1}^{C} \frac{Y_c^R}{p_c} - Y. \tag{2.3}$$

Using the relation $\sum_{hik \in U} \delta_{chik} \left(R_{hik} - p_c \right) = 0,$

$$\begin{aligned} \theta &= \sum_{c=1}^{C} \sum_{hik \in U} y_{hik} \delta_{chik} \left(\frac{R_{hik}}{p_c} - 1 \right) \\ &= \sum_{c=1}^{C} \sum_{hik \in U} \delta_{chik} \left(\frac{R_{hik}}{p_c} - 1 \right) \left(y_{hik} - \frac{Y_c^R}{M_c^R} \right). \end{aligned}$$

We are interested in testing the hypothesis $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0,$ or alternatively in obtaining a confidence interval for θ. If the response propensity in each poststratum c is uniform with $R_{hik} = p_c$ for all units having $\delta_{chik} = 1,$ then θ will be zero. Alternatively, $\theta = 0$ if there is no variability in the response variable y_{hik} within each poststratum. If either of these conditions holds, poststratification corrects for bias from nonresponse. Note that if each of the poststrata has uniform response propensity – that is, the poststratification variables completely explain the variability in underlying response propensities – then the poststratification will in fact remove bias for every possible y variable. If the variance of y_{hik} is 0 within each poststratum, poststratification removes bias for y but it does not necessarily remove bias for other variables.

We estimate θ by $\hat{\theta} = \hat{Y}_{PS} - \hat{Y}_{SS},$ which may be rewritten as

$$\hat{\theta} = \hat{Y}_{PS} - \hat{Y}_{SS} = \sum_{c=1}^{C} \frac{1}{p_c} \left(\hat{Y}_c^R - \bar{Y}_c^R \left(\hat{M}_c^R - M_c^R \right) + \hat{T}_c \right) - \hat{Y}_{SS}, \tag{2.4}$$

where $\bar{Y}_c^R = Y_c^R / M_c^R,$ $\bar{y}_c^R = \hat{Y}_c^R / \hat{M}_c^R,$ and

$$\hat{T}_c = -\left(\bar{y}_c^R - \bar{Y}_c^R\right)\left(\hat{M}_c^R - M_c^R\right). \tag{2.5}$$

Theorem 1 gives the variance of $\hat{\theta}$. Define

$$e_{Rhik} = \sum_{c=1}^{C} \delta_{chik} \left\{ \frac{R_{hik}}{p_c}\left(y_{hik} - \bar{Y}_c^R\right) - y_{hik} \right\}.$$

We assume the following regularity conditions.

(A1) The number of poststrata, C, is fixed and $M_c / M \to \lambda_c \in (0,1)$.

(A2) There exists a constant K such that $|y_{hik}| < K$ for all (hik).

(A3) $\max_{hik} w_{hik} = O(M/n)$ and $\max_{hik} w_{hik}/w_{hi}$ is bounded.

(A4) $R_{hik} > \varepsilon$ for all (hik), for a fixed $\varepsilon > 0$. This guarantees that every unit has a positive response propensity that is bounded away from 0.

(A5) The vector of response indicators $\mathbf{r} = [r_{hik}]$ is independent of the vector of sample inclusion indicators $\mathbf{Z} = [Z_{hik}]$. In addition, r_{hik} and r_{ljp} are independent when $(hi) \neq (lj)$, so that the response indicators in different PSUs are uncorrelated.

Assumptions (A1) and (A4) ensure that the denominator in (2.3) is nonzero almost surely. Assumption (A2) could be replaced by weaker Liapunov-type conditions such as those in Theorem 1.3.2 of Fuller (2009) or Yung and Rao (2000) if more restrictive assumptions are placed on the covariance structure of the response indicators r_{hik}; however, in practice it can be assumed that almost any characteristic measured in a finite population is bounded. Assumption (A5) is weaker than the assumption used in Kim and Kim (2007) that the response indicators are independent across units. With assumption (A5), individuals in the same PSU (for example, persons in the same household or same city) may exhibit dependence when choosing whether to respond to a survey, but the response indicators of individuals in different PSUs are independent.

Theorem 1. *Under conditions (A1) – (A5), the variance of $\hat{\theta}$ is*

$$V\left(\hat{\theta}\right) = V_1\left(\hat{\theta}\right) + V_2\left(\hat{\theta}\right),$$

where

$$V_1\left(\hat{\theta}\right) = V\left(\sum_{hik \in U} Z_{hik} w_{hik} e_{Rhik}\right) + E\left[V\left[\sum_{hik \in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \frac{r_{hik}}{p_c}\left(y_{hik} - \bar{Y}_c^R\right) \middle| \mathbf{Z}\right]\right] \tag{2.6}$$

and

$$V_2\left(\hat{\theta}\right) = V\left[\sum_{c=1}^{C} \frac{\hat{T}_c}{p_c}\right] + 2\ \mathrm{Cov}\left[\sum_{c=1}^{C} \frac{\hat{T}_c}{p_c}, \sum_{c=1}^{C} \frac{\left(\bar{y}_c^R - \bar{Y}_c^R\right)\hat{M}_c^R}{p_c} - \hat{Y}_{SS}\right] = o\left(M^2/n\right).$$

The proof is given in the appendix. Usually, only $V_1\left(\hat{\theta}\right)$ would be considered because for most applications it has higher order than $V_2\left(\hat{\theta}\right)$. Unlike situations typically studied in survey sampling, however, the first-order term of the linearization variance can be zero for some situations, and in those cases

$V\left(\hat{\theta}\right) = V_2\left(\hat{\theta}\right)$. If the first-order term is not exactly zero but has order $o(M^2/n)$, both terms of the variance are needed.

The second term in (2.6) equals 0 if $p_c = 1$ for all poststrata c (that is, there is full response), or if there is no variability among the y values within poststratum c for each poststratum with $p_c < 1$. If the response indicators r_{hik} are all independent, then

$$E\left[V\left(\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \frac{r_{hik}}{p_c}\left(y_{hik} - \bar{Y}_c^R\right) \middle| \mathbf{Z}\right)\right] = \sum_{hik\in U} w_{hik} \sum_{c=1}^{C} \delta_{chik} \frac{1-p_c}{p_c}\left(y_{hik} - \bar{Y}_c^R\right)^2.$$

Under the hypothesized uniform response propensity mechanism that $R_{hik} = p_c$ for all population units in poststratum c, the first term in (2.6) is

$$V\left(\sum_{hik\in U} Z_{hik} w_{hik} e_{Rhik}\right) = V\left\{\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik}\left(-\bar{Y}_c^R\right)\right\} = V\left(\sum_{c=1}^{C} \hat{M}_c \bar{Y}_c^R\right).$$

If response propensities are uniform, this term equals zero if the population mean of $\bar{Y}_c^R$ is the same for all poststrata and the estimated poststratum sizes sum to M.

If $(n/M^2)V_1\left(\hat{\theta}\right)$ converges to a positive constant, a linearization variance estimator for $V\left(\hat{\theta}\right)$ is

$$\hat{V}_L\left(\hat{\theta}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i\in S_h} \left(b_{hi} - b_h\right)^2 \tag{2.7}$$

where

$$b_{hi} = \sum_{k\in S_{hi}} w_{hik} \left\{\sum_{c=1}^{C} \frac{M_c}{\hat{M}_c^R} r_{hik} \delta_{chik}\left(y_{hik} - \bar{y}_c^R\right) - y_{hik}\right\}$$

and

$$b_h = \frac{1}{n_h} \sum_{i\in S_h} b_{hi}.$$

Theorem 2. *Suppose conditions (A1) – (A5) hold and that* $(n/M^2)V_1\left(\hat{\theta}\right)$ *converges to a positive constant. Then* $(n/M^2)\left[\hat{V}_L\left(\hat{\theta}\right) - V_1\left(\hat{\theta}\right)\right] \to 0$ *in probability.*

Theorem 2 is proven in the Appendix.

2.2 Higher-order terms of the variance

When $V_1\left(\hat{\theta}\right) = o(M^2/n)$, the higher-order terms of the variance are needed. Theorem 3 gives these higher-order terms for the special case of simple random sampling. For simple random sampling, each unit is denoted by the subscript i instead of hik.

Theorem 3. *Suppose conditions (A1) – (A5) are met, and that a simple random sample of n units is selected from the population of M units, where $n/M \to 0$. Let $\hat{Y}_c^{NR} = \sum_{i \in S} w_i \delta_{ci} y_i (1 - r_i)$ be the estimated total for the nonrespondents in poststratum c. Assume that $\bar{y}_c^R$ is independent of $\hat{M}_c^R$ and $\hat{Y}_c^{NR}$, and that all r_i are independent and are independent of Z_i. Then*

$$V_2\left(\hat{\theta}\right) = \sum_{c=1}^{C} \frac{2p_c - 1}{p_c^2} V\left[\bar{y}_c^R - \bar{Y}_c^R\right] V\left[\hat{M}_c^R - M_c^R\right] + o(M^2/n^2).$$

We can estimate $V_2\left(\hat{\theta}\right)$ in a simple random sample by

$$\sum_{c=1}^{C} \frac{2\hat{p}_c - 1}{\hat{p}_c^2} \frac{s_c^2}{n_c^R} \frac{M_c \hat{p}_c \left(M - M_c \hat{p}_c\right)}{n},$$

where $\hat{p}_c$ is the empirical response rate in poststratum c, n_c^R is the number of respondents in poststratum c, and s_c^2 is the sample variance of y for the respondents in poststratum c.

In practice, the estimated first-order term of the variance using (2.7) will in general be nonzero even when $V_1\left(\hat{\theta}\right) = 0$. Thus, the estimated first-order term cannot be used to diagnose whether the higher-order terms are needed. However, the variance expression in (2.6) implies that $V_1\left(\hat{\theta}\right)$ is sufficiently large for the first-order approximation to be valid when all poststrata have response rates bounded away from one and non-negligible within-poststratum variance.

2.3 Jackknife

The jackknife estimator of the variance is defined as follows:

$$\hat{V}_J\left(\hat{\theta}\right) = \sum_{g=1}^{H} \frac{n_g - 1}{n_g} \sum_{j \in S_g} \left(\hat{\theta}^{(gj)} - \hat{\theta}\right)^2, \tag{2.8}$$

where

$$\begin{aligned}
\hat{\theta}^{(gj)} &= \hat{Y}_{PS}^{(gj)} - \hat{Y}_{SS}^{(gj)}, \\
\hat{Y}_{PS}^{(gj)} &= \sum_{c=1}^{C} M_c \frac{\sum_{hik \in S} w_{hik}^{(gj)} r_{hik} \delta_{chik} y_{hik}}{\sum_{hik \in S} w_{hik}^{(gj)} r_{hik} \delta_{chik}}, \\
\hat{Y}_{SS}^{(gj)} &= \sum_{hik \in S} w_{hik}^{(gj)} y_{hik},
\end{aligned}$$

and the jackknife weights are:

$$w_{hik}^{(gj)} = \begin{cases} 0 & \text{if} \quad (hi) = (gj) \\ \dfrac{n_h}{n_h - 1} w_{hik} & \text{if} \quad h = g, i \neq j \\ w_{hik} & \text{if} \quad h \neq g \end{cases}. \tag{2.9}$$

If $(n/M^2)V_1(\hat{\theta})$ converges to a positive constant and assumptions (A1) – (A5) hold, then $\hat{V}_J(\hat{\theta})/V_1(\hat{\theta})$ converges to 1 in probability. This follows by standard jackknife arguments (Theorem 6.1 of Shao and Tu 1995) because the population parameter is a continuously differentiable function of population totals. Under the conditions of Theorem 2, either $\hat{\theta}/\sqrt{\hat{V}_L(\hat{\theta})}$ or $\hat{\theta}/\sqrt{\hat{V}_J(\hat{\theta})}$ may be used as a test statistic. Each approximately follows a standard normal distribution when the null hypothesis $H_0: \theta = 0$ is true.

2.4 Remarks and extensions

In this section we derived the linearization variance estimator for comparing the estimated population total of a quantity known for everyone in the selected sample with the poststratified estimate calculated using the respondents only. Theorems 1 and 2 also give the variance and variance estimator for comparing the estimator calculated using the selected sample with that from the base-weighted respondents. In that case, $\hat{Y}_{PS}$ reduces to an estimator with one poststratum, $\hat{Y}_{PS} = \left(M/\hat{M}^R\right)\hat{Y}^R$, where $\hat{M}^R = \sum_{(hik)\in S} w_{hik} r_{hik}$.

What happens if y is one of the poststratification variables? In the framework used in this section, the population counts for the poststratification variables are obtained from the sampling frame or an external source. If y is a linear combination of poststratification class indicators, then $\hat{Y}_{PS}$ is the same for all possible samples and thus has zero variance. Then $V(\hat{\theta}) = V(\hat{Y}_{SS})$, which is the first-order term of the variance in Theorem 1. If y is also a stratification variable in the design, then $V(\hat{\theta})$ will be zero. If y is not a stratification variable, then typically $\hat{Y}_{SS}$ will vary from sample to sample and will have variance of order $O(M^2/n)$ so that the test of nonresponse bias can be performed. We would expect the rejection rate for the test to be the significance level α in this case.

The parameter θ in (2.3) was defined as the difference between the poststratified population total, calculated using the population response propensities under the poststratification scheme adopted, and the unadjusted population total. In (2.4), the unadjusted population total Y was estimated by the Horvitz-Thompson estimator. The parameter θ could alternatively be estimated by

$$\hat{\theta}_2 = \hat{Y}_{PS} - \sum_{c=1}^{C} M_c \frac{\hat{Y}_c}{\hat{M}_c},$$

in which a poststratified estimator is used instead of $\hat{Y}_{SS}$. The variance of $\hat{\theta}_2$ is expected to be less than the variance of $\hat{\theta}$ under the poststratification assumptions, resulting in a more powerful test. However, when y is a linear combination of the poststratum indicators, the statistic $\hat{\theta}_2$ cannot be used to test $H_0: \theta = 0$ because $V(\hat{\theta}_2) = 0$. A similar problem can occur when y is highly correlated with the poststratification variables. The estimator $\hat{\theta}$, by contrast, typically has positive variance even when y is one of the poststratification variables.

Sometimes poststratification is performed using less-than-perfect poststratification totals – for example, the totals may come from a large survey such as the American Community Survey which has its own sampling and nonsampling errors, or they may be from a census of a slightly different population. In some cases, poststratification variables such as race or ethnicity may be measured differently in the survey than in the source of the external population totals. Using $\hat{\theta}$ rather than $\hat{\theta}_2$ may detect differences that might be caused by a flawed poststratification.

If desired, the tests may be performed using means rather than totals. In this case, the population parameter is

$$\theta_M = \sum_{c=1}^{C} \frac{M_c}{M} \frac{Y_c^R}{M_c^R} - \bar{Y}$$

where $\bar{Y} = Y/M$, and may be estimated by

$$\hat{\theta}_M = \sum_{c=1}^{C} \frac{M_c}{M} \frac{\hat{Y}_c^R}{\hat{M}_c^R} - \frac{\hat{Y}_{SS}}{\sum_{hik \in S} w_{hik}}. \tag{2.10}$$

3 Propensity weighting

An alternative to poststratification is to use inverse propensity weighting of the respondents (see, for example, Folsom 1991; Kim and Kim 2007).

In this framework, the true response propensity of unit (hik) is R_{hik} and a model is used to predict the propensity from characteristics known for everyone in the selected sample. Logistic regression is often used to estimate propensities. Suppose that the p- vector $\mathbf{x}_{hik}$ is known for each unit in S. The modeled response propensity, if $\mathbf{x}_{hik}$ and R_{hik} were known for each unit in the population, is

$$R_{hik}^M = \left[1 + \exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\right]^{-1},$$

where $\boldsymbol{\beta}$ is the solution to the expected population score equations

$$\sum_{hik \in U} \left[R_{hik} - R_{hik}^M\right]\mathbf{x}_{hik} = 0.$$

The model removes the bias for the estimated population total of y if

$$\theta = \sum_{hik \in U} \left[R_{hik} \frac{y_{hik}}{R_{hik}^M} - y_{hik}\right] \tag{3.1}$$

equals 0. If $R_{hik} = R_{hik}^M$, that is, the response propensity model is correctly specified, then the weighting adjustments remove the bias for every possible response variable y. The population parameter θ is estimated by

$$\hat{\theta} = \sum_{hik \in S} w_{hik}\left[r_{hik} y_{hik}\left[1 + \exp\left(-\mathbf{x}_{hik}'\hat{\boldsymbol{\beta}}\right)\right] - y_{hik}\right],$$

where $\hat{\boldsymbol{\beta}}$ is the solution to the pseudolikelihood score equations

$$\sum_{hik \in S} w_{hik}\left[r_{hik} - \left[1 + \exp\left(-\mathbf{x}_{hik}'\hat{\boldsymbol{\beta}}\right)\right]^{-1}\right]\mathbf{x}_{hik} = 0.$$

Unlike the poststratification situation, the population parameter θ in (3.1) is not an explicit function of population totals. Similarly to Kim and Kim (2007), we can obtain the linearization variance and a linearization variance estimator of $\hat{\theta}$ by using the estimating equation for $(\boldsymbol{\beta}, \theta)$, as derived in Binder (1983): $\left(\hat{\boldsymbol{\beta}}, \hat{\theta}\right)$ is the solution to

$$\hat{\mathbf{A}}(\boldsymbol{\beta},\theta,\mathbf{r}) = \sum_{hik\in S} w_{hik}\mathbf{u}\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right) - [0,0,\ldots,0,\theta]' = 0, \tag{3.2}$$

where

$$\mathbf{u}\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right) = \begin{bmatrix}\mathbf{u}_1\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right)\\ u_2\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right)\end{bmatrix} = \begin{bmatrix}\left[r_{hik} - \left[1+\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\right]^{-1}\right]\mathbf{x}_{hik}\\ r_{hik}y_{hik}\left[1+\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\right] - y_{hik}\end{bmatrix}.$$

The population parameter θ solves the population estimating equation

$$\mathbf{A}(\boldsymbol{\beta},\theta,\mathbf{R}) = \sum_{hik\in U}\mathbf{u}\left(y_{hik},\mathbf{x}_{hik},R_{hik},\boldsymbol{\beta}\right) - [0,0,\ldots,0,\theta]' = 0.$$

Theorem 4. *Let* $\hat{\mathbf{U}}(\boldsymbol{\beta},\theta) = \sum_{hik\in S} w_{hik}\mathbf{u}\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right) = \left[\hat{\mathbf{U}}_1(\boldsymbol{\beta})', \hat{U}_2(\boldsymbol{\beta})\right]'$. *Suppose conditions (A2) – (A5) are met and there exists a value* B *such that* $\left|\mathbf{x}_{hik,j}\right| < B$ *for all units* (hik) *and components* j. *Then* $V\left(\hat{\theta}\right) = V_L\left(\hat{\theta}\right) + o(M^2/n)$, *where*

$$V_L\left(\hat{\theta}\right) = \mathbf{T}'\mathbf{Q}\mathbf{X}\mathbf{C}V\left[\hat{\mathbf{U}}_1(\boldsymbol{\beta})\right]\mathbf{C}\mathbf{X}'\mathbf{Q}\mathbf{T} - 2\mathbf{T}'\mathbf{Q}\mathbf{X}\mathbf{C}Cov\left[\hat{\mathbf{U}}_1(\boldsymbol{\beta}),\hat{U}_2(\boldsymbol{\beta})\right] + V\left[\hat{U}_2(\boldsymbol{\beta})\right], \tag{3.3}$$

$\mathbf{X}$ *is the* $M\times p$ *matrix with rows* $\mathbf{x}_{hik}'$, $\mathbf{T}$ *is the* M*-vector with elements* $R_{hik}y_{hik}$, $\mathbf{Q}$ *is the* $M\times M$ *diagonal matrix with entries* $\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)$, *and* $\mathbf{C} = \left(\mathbf{X}'[\mathbf{I}+\mathbf{Q}]^{-2}\mathbf{Q}\mathbf{X}\right)^{-1}$.

A linearization variance estimator for $\hat{\theta}$ may be obtained by substituting estimators for the population quantities in (3.3) to obtain

$$\begin{aligned}\hat{V}_L\left(\hat{\boldsymbol{\beta}},\hat{\theta}\right) &= \mathbf{t}_S'\mathbf{W}_S\mathbf{Q}_S\mathbf{X}_S\hat{\mathbf{C}}\hat{V}\left[\hat{\mathbf{U}}_1\left(\hat{\boldsymbol{\beta}}\right)\right]\hat{\mathbf{C}}\mathbf{X}_S'\mathbf{Q}_S\mathbf{W}_S\mathbf{t}_S\\ &\quad - 2\mathbf{t}_S'\mathbf{W}_S\mathbf{Q}_S\mathbf{X}_S\hat{\mathbf{C}}\widehat{\mathrm{Cov}}\left[\hat{\mathbf{U}}_1\left(\hat{\boldsymbol{\beta}}\right),\hat{U}_2\left(\hat{\boldsymbol{\beta}}\right)\right] + \hat{V}\left[\hat{U}_2\left(\hat{\boldsymbol{\beta}}\right)\right],\end{aligned}$$

where $\mathbf{X}_S$ is the $m\times p$ matrix with rows $\mathbf{x}_{hik}'$ for the sampled units with m the size of the selected sample, $\mathbf{W}_S$ is the $m\times m$ diagonal matrix of weights w_{hik} for sampled units, $\mathbf{t}_S$ is the m-vector with elements $r_{hik}y_{hik}$ for sampled units, $\mathbf{Q}_S$ is the $m\times m$ diagonal matrix with entries $\exp\left(-\mathbf{x}_{hik}'\hat{\boldsymbol{\beta}}\right)$ for values of $\mathbf{x}_{hik}$ in the sample, and $\hat{\mathbf{C}} = \left(\mathbf{X}_S'\mathbf{W}_S\left[\mathbf{I}+\mathbf{Q}_S\right]^{-2}\mathbf{Q}_S\mathbf{X}_S\right)^{-1}$.

The jackknife variance estimator for inverse propensity weighting is defined using the formula in (2.8) with jackknife weights in (2.9). For the propensity setting,

$$\hat{\theta}^{(gj)} = \sum_{hik\in S} w_{hik}^{(gj)}\left[r_{hik}y_{hik}\left[1+\exp\left(-\mathbf{x}_{hik}'\hat{\boldsymbol{\beta}}^{(gj)}\right)\right] - y_{hik}\right],$$

where $\hat{\boldsymbol{\beta}}^{(gj)}$ solves

$$\sum_{hik\in S} w_{hik}^{(gj)}\left[r_{hik} - \left[1+\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\right]^{-1}\right]\mathbf{x}_{hik} = 0.$$

Theorem 5. *Assume that the conditions of Theorem 4 hold. If* $(n/M^2)V_L\left(\hat{\theta}\right)$ *converges to a positive constant, then* $(n/M^2)\left[\hat{V}_L\left(\hat{\theta}\right) - V_L\left(\hat{\theta}\right)\right]$ *and* $(n/M^2)\left[\hat{V}_J\left(\hat{\theta}\right) - V_L\left(\hat{\theta}\right)\right]$ *both converge in probability to 0.*

The proof of Theorem 5 follows by standard arguments in Fuller (2009) and Shao and Tu (1995) and is hence omitted.

The parameter θ for examining bias with inverse propensity weighting was defined for population totals. As with poststratification, it may be desired to compare means instead of totals, particularly if weight trimming is used to truncate large and influential values of the propensity weight $\left[1+\exp\left(-\mathbf{x}'_{hik}\hat{\boldsymbol{\beta}}\right)\right]$. In this case, the parameter to be evaluated is

$$\theta_M = \frac{\sum_{hik\in U}\left[R_{hik}y_{hik}/R^M_{hik}\right]}{\sum_{hik\in U}R_{hik}/R^M_{hik}} - \frac{\sum_{hik\in U}y_{hik}}{M}$$

with estimator

$$\hat{\theta}_M = \frac{\sum_{hik\in S}r_{hik}w_{hik}y_{hik}\left[1+\exp\left(-\mathbf{x}'_{hik}\hat{\boldsymbol{\beta}}\right)\right]}{\sum_{hik\in S}r_{hik}w_{hik}\left[1+\exp\left(-\mathbf{x}'_{hik}\hat{\boldsymbol{\beta}}\right)\right]} - \frac{\sum_{hik\in S}w_{hik}y_{hik}}{\sum_{hik\in S}w_{hik}}.$$

Special adjustments are needed to account for weight trimming with the linearization variance estimator; in general, we recommend using the jackknife or another replication method for finding the variance of $\hat{\theta}$ or $\hat{\theta}_M$.

4 Simulation results

We examine the performance of the variance estimators in two simulation studies. The first study generates finite populations with response indicators r_{hik} and then draws simple random samples from the population. The second simulation uses data from the 2009-2013 5-year American Community Survey Public Use Microdata Samples (ACS PUMS) as a population and then draws repeated cluster samples from this population under different nonresponse mechanisms.

For the simulation involving simple random sampling, we generated finite populations of 1,000,000 units. To study the poststratification estimator we used $C = 6$ poststrata to generate nonresponse. The experimental factors were:

- sample size, n: 300 or 1,000.
- population proportion $\left(M_c/M\right)$ in each poststratum: (P1) (1/6, 1/6, 1/6, 1/6, 1/6, 1/6), (P2) (1/21, 2/21, 3/21, 4/21, 5/21, 6/21), and (P3) (6/21, 5/21, 4/21, 3/21, 2/21, 1/21).
- response rates in poststrata: (R1) (0.2, 0.3, 0.5, 0.6, 0.8, 0.9), (R2) (0.3, 0.7, 0.3, 0.7, 0.3, 0.7), and (R3) (1, 1, 1, 1, 1, 1). Level (R3), with full response, is included to explore the accuracy of the higher-order approximation to the variance when $V_1\left(\hat{\theta}\right) = 0$.
- poststratum means: (M1) (0, 0, 0, 0, 0, 0), (M2) (-2, -1, 0, 1, 2, 3) and (M3) (0, 1, 0, 1, 0, 1).
- number of poststrata used in nonresponse adjustment: 1, 3 (collapse adjacent pairs of poststrata), or 6. Only the settings with 6 poststrata are guaranteed to correct for the nonresponse bias.

Within each poststratum, population values y_i were generated from a normal distribution with the specified poststratum mean and variance 1. The response indicators r_i were generated as independent

Bernoulli random variables with mean R_i. The simple random sampling simulations were done in version 3.2.2 of R (R Core Team 2015), and 2,000 iterations were performed for each of the 162 simulation settings, which results in a standard error less than 0.005 for the Monte Carlo estimate of the rejection proportion when the null hypothesis of $\theta = 0$ is true. Some of the generated samples had fewer than two respondents in one or more poststrata, which would result in some jackknife resamples having no respondents in those poststrata. For such samples, the two poststrata with the smallest number of respondents were combined iteratively until all poststrata had at least two respondents.

For each simulation setting, the Monte Carlo (MC) variance of $\hat{\theta}$, $\hat{V}_{MC}\left(\hat{\theta}\right)$, was calculated as the sample variance of $\hat{\theta}_b$ for $b = 1,\ldots, 2{,}000$. The linearization and jackknife variance estimates were calculated for each simulated sample, and the means of those estimates over the 2,000 samples are denoted as $\hat{V}_L\left(\hat{\theta}\right)$ and $\hat{V}_J\left(\hat{\theta}\right)$, respectively.

Figures 4.1 and 4.2 display results for the simulation settings in which $V_1\left(\hat{\theta}\right) > 0$. Figure 4.1 displays histograms of the ratios of the mean linearization and jackknife variance estimates to $\hat{V}_{MC}\left(\hat{\theta}\right)$. The scatterplot in Figure 4.2 displays the percentage of the 2,000 iterations in which the null hypothesis $H_0 : \theta = 0$ is rejected at the 5% significance level. Most of the variance estimates are close to the MC variance and the rejection rate for $H_0 : \theta = 0$ is approximately 5% when $\theta = 0$, with higher power for larger values of $|\theta|$. Four of the simulation runs with $\theta = 0$, however, have linearization and jackknife variances that are approximately twice the MC variance, and rejection rates that are between 0 and 1%. These results are from the simulations with poststratum means (M3), response rates (R3), population proportions (P2) or (P3), and three collapsed poststrata. Although the population means for the collapsed poststrata differ, they do not differ greatly and a sample size of 1,000 is too small for the first-order asymptotic approximation to be accurate. For these settings, a sample size of approximately 15,000 was needed to reduce the variance ratios $\hat{V}_L\left(\hat{\theta}\right)/\hat{V}_{MC}\left(\hat{\theta}\right)$ and $\hat{V}_J\left(\hat{\theta}\right)/\hat{V}_{MC}\left(\hat{\theta}\right)$ to 1.2.

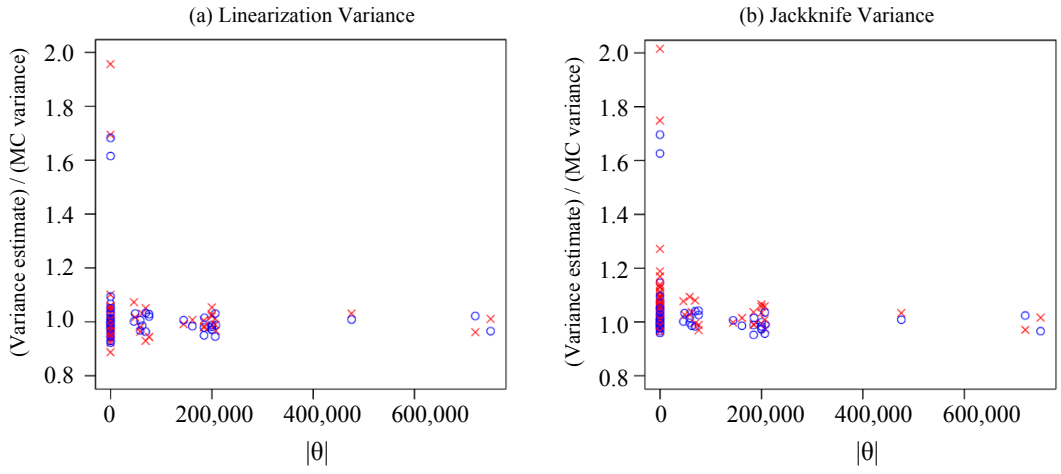

Figure 4.1 Ratios of (a) $\hat{V}_L\left(\hat{\theta}\right)$ and (b) $\hat{V}_J\left(\hat{\theta}\right)$ to $\hat{V}_{MC}\left(\hat{\theta}\right)$, for the simple random sample poststratification simulation settings in which $V_1\left(\hat{\theta}\right) > 0$. The blue circles represent simulations with $n = 1{,}000$ and the red Xs represent simulations with $n = 300$.

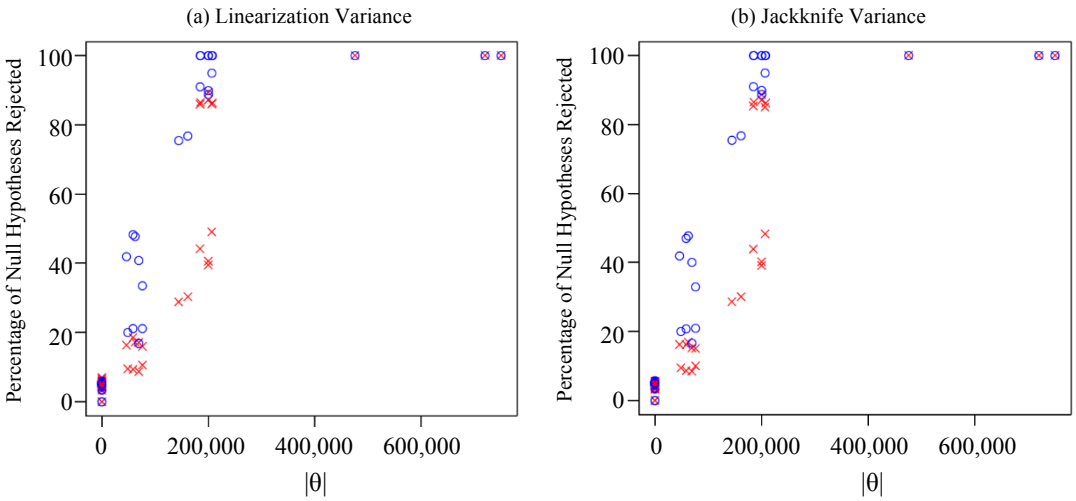

Figure 4.2 Empirical power for tests using linearization and jackknife variance, for the simple random sample poststratification simulation settings in which $V_1(\hat{\theta}) > 0$. The blue circles represent simulations with $n = 1{,}000$ and the red Xs represent simulations with $n = 300$.

Figure 4.3 shows the behavior of $\hat{V}_L(\hat{\theta})$, $\hat{V}_J(\hat{\theta})$, and $\hat{V}_2(\hat{\theta})$ when the first-order term of the variance is $V_1(\hat{\theta}) = 0$ but $V_2(\hat{\theta}) > 0$. For all of those simulations, the true value of θ was 0 and the second-order term $\hat{V}_2(\hat{\theta})$ was calculated using the SRS approximation in Theorem 3. Even though the true first-order variance $V_1(\hat{\theta})$ is zero for these settings, the estimated first-order variances from linearization and jackknife are nonzero. For the simulations with poststratum means (M1) and response rates (R3), for example, all poststrata have the same population mean. The sample means for the poststrata differ, however, and this causes the linearization and jackknife variance estimators to be positive and, on average, about twice as large as the MC variance. The same thing happens with poststratum means (M3), population proportions (P1), and response rates (R3) when three poststrata are used: the three collapsed poststrata each have population mean 1/2 but the sample means vary.

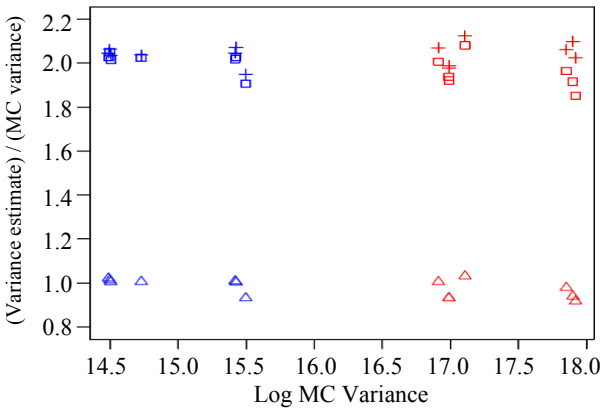

Figure 4.3 Ratios of $\hat{V}_L(\hat{\theta})$ (squares), $\hat{V}_J(\hat{\theta})$ (plus signs), and $\hat{V}_2(\hat{\theta})$ (triangles) to $\hat{V}_{MC}(\hat{\theta})$, plotted against $\ln\hat{V}_{MC}(\hat{\theta})$, for the simple random sample poststratification simulation settings in which $V_1(\hat{\theta}) = 0$. For all of these settings, $\theta = 0$. The blue symbols (with log MC variance < 16) represent simulations with $n = 1{,}000$ and the red symbols (with log MC variance > 16) represent simulations with $n = 300$.

Only simulation settings with response rates (R3) required the use of higher-order terms or large sample sizes for the linearization and jackknife variance estimators to be accurate. It would be easy to identify these situations in practice from the absence of nonresponse.

To study the properties of the estimators in Section 3, we used a subset of the populations generated for the poststratification simulation as well as populations generated with continuous covariate x, giving factors:

- Sample size, n: 300 or 1,000.
- Population values and nonresponse generation.
  1. Nonresponse is generated in 6 poststrata with population proportions (P1) or (P2), and response rates (R1) or (R2). The variable of interest y is generated with poststratum means (M1) and (M2) plus a $N(0,1)$ error term.
  2. Covariate x is generated from a $N(0,1)$ distribution. Then y is generated as (Y1) $0+N(0,1)$ (independent of x), (Y2) $x+N(0,1)$, or (Y3) $x^2+N(0,1)$. The response propensities are generated as (R1P) $R=0.8$ for all units, (R2P) $\text{logit}(R)=1/(1+\exp(-x))$, and (R3P) $\text{logit}(R)=1/(1+\exp(-x^2/3))$.
- Response propensity model used.
  1. For poststratified populations, treat x as a continuous variable with values 1–6.
  2. For populations with generated covariate x, use linear logistic regression with covariate x. This model is correctly specified for response-generating mechanisms (R1P) and (R2P) but incorrectly specified for mechanism (R3P).

To reduce the instability of the estimators, estimated response propensities less than 0.05 were replaced by 0.05, corresponding to trimming weight adjustments larger than 20. Figures 4.4 and 4.5 display the variance ratios and empirical power for the propensity model simulations. All settings in this simulation had $V_1\left(\hat{\theta}\right)>0$. As in the poststratification simulation, the linearization and jackknife variance estimators both perform well in general. There are a few settings, however, in which the linearization variance is substantially larger than the jackknife. This occurs because of the weight trimming: the jackknife automatically accounts for the effect of weight trimming on the variance because the jackknife replicates also trim the weights. The linearization variance used in this simulation was from Theorem 5, and the formula would need to be modified to include the effects of trimming. We also ran simulations using the jackknife in which the mean was estimated instead of the population total, and the jackknife performed well for that parameter as well.

The second simulation study used a population of 6,019,599 household-level records from the ACS PUMS studied in Lohr, Hsu and Montaquila (2015). There are 3,344 PSUs in the population defined by the public use microdata areas. Eight poststrata were formed based on the cross-classification of households by tenure (rent or own), presence of children in the household (yes or no), and number of income earners (0-1 or 2+). The primary outcome variable y was household income. Additionally, a less skewed outcome variable $\log(y)$ was studied, where $\log(y)$ was set to 0 if $y<1$.

A $2 \times 2 \times 3$ factorial design was used for this study with factors

- overall response rate: 50% or 80%.
- number of PSUs for each sample: 25 or 100.
- nonresponse generating mechanism: (N1) missing completely at random (MCAR), with response propensity for all records equal to the response rate for all households; (N2) missing at random (MAR), where a linear logistic model with main effect terms for tenure, presence of children, and number of income earners generates the response propensities; and (N3) missing not at random (MNAR), where a linear logistic model with main effect terms for tenure, presence of children, and household income generates the response propensities.

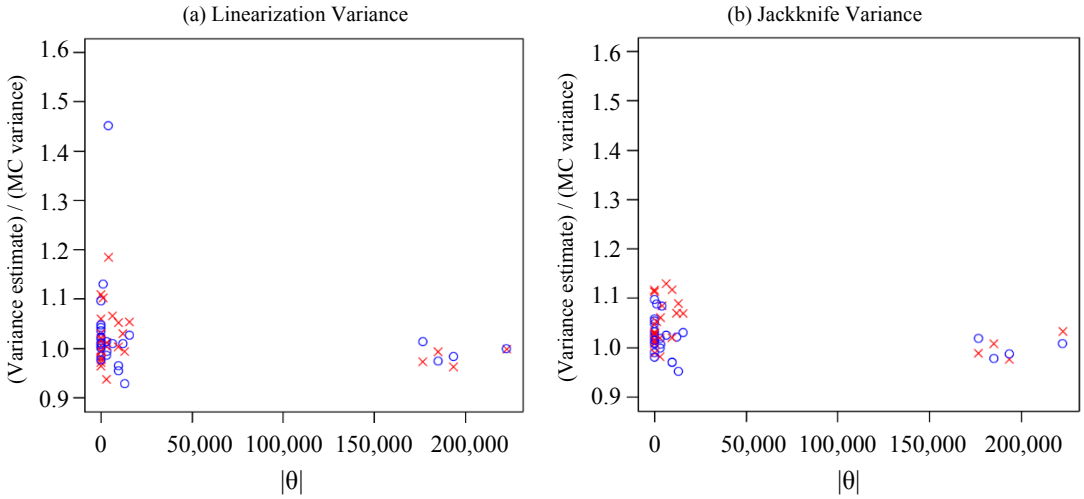

Figure 4.4 Ratios of (a) $\hat{V}_L(\hat{\theta})$ and (b) $\hat{V}_J(\hat{\theta})$ to $\hat{V}_{MC}(\hat{\theta})$, for the propensity model simulation. The blue circles represent simulations with $n = 1{,}000$ and the red Xs represent simulations with $n = 300$.

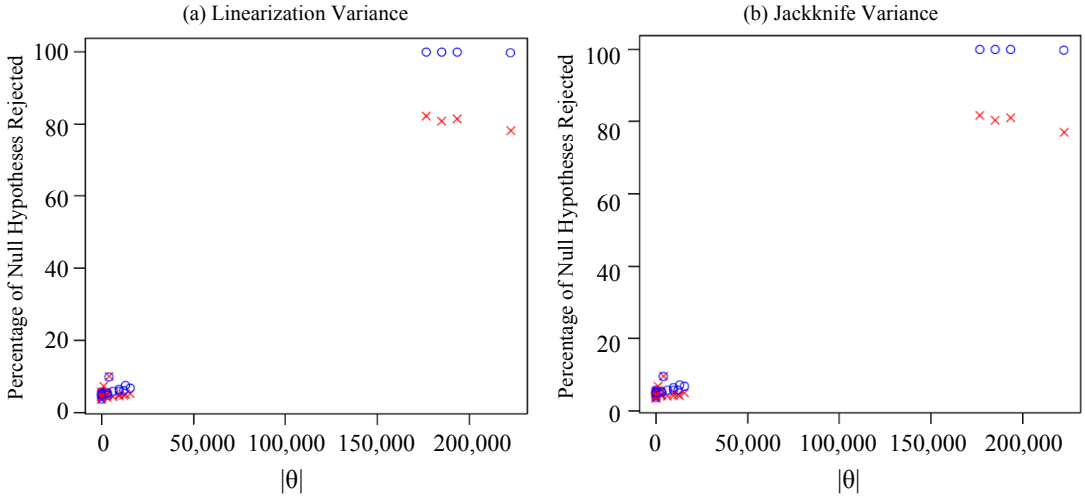

Figure 4.5 Empirical power for tests in the propensity model simulation using linearization and jackknife variance. The blue circles represent simulations with $n = 1{,}000$ and the red Xs represent simulations with $n = 300$.

For the first two nonresponse generating mechanisms, $\theta = 0$. For the first mechanism, there is no nonresponse bias. Poststratification corrects for the bias in the second mechanism because $R_{hik} = p_c$ for units in poststratum c. Poststratification does not correct for the bias in the third mechanism because the nonresponse depends on the y variable, household income.

For each simulation setting, response indicators were generated independently for the population units using the calculated response propensities. One thousand samples were drawn for each setting, in which PSUs were selected with probability proportional to size and a simple random sample of 100 households was selected from each sampled PSU. The standard error for the rejection proportion when $\theta = 0$ is less than 0.007.

Calculations for the ACS simulation were done in SAS® software (SAS Institute, Inc. 2011). We first calculated the weights and jackknife weights for the selected sample, and then calculated the poststratified and jackknife poststratified weights for the respondents. The two sets of jackknife weights used the same replication structure, so that replicate weight k for the respondents deleted the same PSU as replicate weight k for the selected sample. To simplify computation of $\hat{\theta}_M$ in (2.10), we concatenated the selected sample and respondents, with their respective weights, into one data set and set $u_i = 1$ for records in the respondent data set and $u_i = 0$ for records in the selected sample data set. The linear model $y_i = \beta_0 + \beta_1 u_i$ was fit to the concatenated data using the SURVEYREG procedure, and $\hat{\theta}_M = \hat{\beta}_1$ from the regression model.

Table 4.1 gives the results from the simulation. For all but one of the simulation settings, the mean of the jackknife variance estimates is larger than the Monte Carlo variance of $\hat{\theta}_M$, but the bias of the jackknife variance is reduced when more PSUs are sampled or the response rate is higher. The outcome variable y, household income, is highly skewed, and the rejection rate when $\theta_M = 0$ is closer to the nominal α of 0.05 when the log-transformed variable is used.

Table 4.1
Simulation results from ACS population

Nonresponse Mechanism	Response Rate (%)	Number of PSUs	Outcome variable y			Outcome variable $\log(y)$		
			θ_M	% Reject	$\frac{\hat{V}_J(\hat{\theta}_M)}{\hat{V}_{MC}(\hat{\theta}_M)}$	θ_M	% Reject	$\frac{\hat{V}_J(\hat{\theta}_M)}{\hat{V}_{MC}(\hat{\theta}_M)}$
MCAR	50	25	0	3.3	1.21	0	4.5	1.20
MCAR	50	100	0	3.0	1.09	0	4.4	1.08
MCAR	80	25	0	3.8	1.14	0	4.0	1.19
MCAR	80	100	0	3.9	1.07	0	5.2	1.05
MAR	50	25	0	4.5	1.16	0	4.2	1.11
MAR	50	100	0	4.9	1.04	0	4.4	1.05
MAR	80	25	0	3.5	1.16	0	4.7	1.20
MAR	80	100	0	3.5	1.12	0	4.6	1.11
NMAR	50	25	8,882	70.8	1.41	0.118	6.3	1.60
NMAR	50	100	8,882	99.5	1.11	0.118	37.7	1.11
NMAR	80	25	3,706	45.6	1.18	0.047	14.5	1.20
NMAR	80	100	3,706	99.4	1.09	0.047	61.0	0.99

5 Discussion

In this paper, we considered tests for nonresponse bias after poststratification or inverse propensity weighting has been used. The arguments in the theorems could be extended to similar methods that are used to adjust for nonresponse bias such as raking, which iteratively poststratifies to marginal population totals, or calibration, which adjusts the weights so that estimated population totals agree with control totals for a set of auxiliary variables. Haziza and Lesage (2016) argued that using a two-step procedure of propensity weighting followed by calibration provides more protection against nonresponse bias than using calibration alone in a single step, because single-step calibration implies a model relating the response propensities and the calibration variables and that model may be misspecified. The tests proposed in this paper could be extended to situations in which both propensity weighting and poststratification are used, or could be used separately to assess the bias removed in each step of a two-step process.

We employed the jackknife for the replication variance estimation. However, all of the estimators are smooth functions of population totals, so other replication variance estimators such as balanced repeated replication or bootstrap could be used as well.

A challenge for evaluating nonresponse bias is the limited amount of information available for the selected sample. For some surveys all available auxiliary information is used or considered for forming poststrata, raking classes, or inverse propensity weights. The poststratified estimator for characteristics used in the poststratification has no variance or bias, so testing these or closely related characteristics will not uncover nonresponse bias in other survey variables. Auxiliary variables that are not used for nonresponse adjustments are often omitted only because they were not selected in model selection method used to form the poststrata or select variables for the logistic regression, and that typically occurs because they have low explanatory power for predicting the response indicator after the other variables are included in the model. For surveys with less frame information, it may be possible to obtain auxiliary information from other sources, such as administrative records associated with the respondents' addresses or paradata. It is important to make sure that the variables used to test nonresponse bias are recorded consistently for respondents and nonrespondents. If, for example, y is the interviewer's curbside assessment about whether children are present in the household, that initial assessment should be used for both respondents and nonrespondents: the assessment used in the nonresponse bias analysis should not be updated after the interviewer ascertains the actual number of children in a responding household.

After testing available variables for nonresponse bias, we still do not know whether the adjustments have removed the bias for outcome variables that are available only for the respondents. Abraham, Helms and Presser (2009) and Kohut, Keeter, Doherty, Dimock and Christian (2012) found that estimates of volunteering and civic participation are higher from surveys with low response rates than from the Current Population Survey, indicating that weighting adjustments do not remove bias for civic engagement variables although they appear to remove bias for demographic variables and home ownership. But testing a wide range of auxiliary variables for residual bias may give more confidence in the results of a survey on the untested variables, or may indicate concerns about inferences from the survey for variables of interest. We recommend that survey designers plan the survey with nonresponse bias assessment in mind, and collect

additional information for the selected sample whenever possible. In general, the more information that can be collected about the selected sample, the better.

The comparison of estimates using different sets of weights may be of special interest when studying responsive or adaptive design strategies such as those described in Groves and Heeringa (2006) and summarized in Tourangeau, Brick, Lohr and Li (2016). In these, later phases of the design are modified using information gleaned in the early returns. One responsive design strategy may be to estimate response rates after the first phase of the survey, and then to allocate resources in the second phase to equalize rates across subgroups of interest. In an experimental comparison of different responsive design strategies, it may be of interest to evaluate the estimated nonresponse bias from the strategies. Riddles, Marker, Rizzo, Wiley and Zukerberg (2015) compared nonresponse-weighted estimates from different data cutoff points in the U.S. Schools and Staffing Survey, to see if estimates changed with earlier truncation of data collection.

The results in Theorems 1 through 5 are expressed for probability samples. There is increased interest in using nonprobability samples to study populations (Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau 2013). Proponents of nonprobability samples argue that with response rates sometimes below 10%, an inexpensive large nonprobability sample can have smaller mean squared error than a small probability sample. The same methods of poststratification and inverse propensity weighting are typically used with nonprobability samples. The tests proposed in this paper can be adapted for use with nonprobability samples, provided that auxiliary information is known for a collection of individuals that can serve as a stand-in for a sampling frame. For a web survey, it might be possible to compare characteristics of persons visiting the web page with those of persons completing the survey. Further research is needed in this area.

Acknowledgements

The authors thank the reviewers for their helpful suggestions that led to improvements in the article.

Appendix

The following lemma shows that the additional variability due to the stochastic response mechanism is $O(M^2/n)$.

Lemma 1. *Suppose assumptions (A3) and (A5) are met, and that* $|q_{hik}| \le Q$ *for all* $(hik) \in U$. *Then*

$$E\left[V\left(\sum_{hik\in U} Z_{hik} w_{hik} q_{hik} r_{hik} \middle| \mathbf{Z}\right)\right] = O(M^2/n).$$

Proof. By assumption (A5),

$$\left| E\left[V\left(\sum_{hik\in U} Z_{hik} w_{hik} q_{hik} r_{hik} \middle| \mathbf{Z} \right) \right] \right| = \left| E\left[\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}\sum_{p=1}^{M_{hi}} Z_{hik} Z_{hip} w_{hik} w_{hip} \mathrm{Cov}\left(r_{hik}, r_{hip}\right) q_{hik} q_{hip} \right] \right|$$

$$\leq Q^2 E\left[\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}\sum_{p=1}^{M_{hi}} Z_{hik} Z_{hip} w_{hik} w_{hip} \right]$$

$$= Q^2 \sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}\sum_{p=1}^{M_{hi}} P\left[(hi)\in S\right] P\left[k\in S_{hi}, p\in S_{hi}\right] w_{hik} w_{hip}$$

$$= O\left(M^2/n\right).$$

The last line is implied by (A3).

Proof of Theorem 1. From (2.4),

$$V_1\left(\hat{\theta}\right) = V\left[\sum_{c=1}^{C} \frac{1}{p_c} \left(\hat{Y}_c^R - \bar{Y}_c^R \left(\hat{M}_c^R - M_c^R \right) \right) - \hat{Y}_{SS} \right]$$

and

$$V_2\left(\hat{\theta}\right) = V\left[\sum_{c=1}^{C} \frac{\hat{T}_c}{p_c} \right] + 2\ \mathrm{Cov}\left[\sum_{c=1}^{C} \frac{\hat{T}_c}{p_c}, \sum_{c=1}^{C} \frac{\left(\bar{y}_c^R - \bar{Y}_c^R\right) \hat{M}_c^R}{p_c} - \hat{Y}_{SS} \right].$$

The leading term simplifies to

$$V_1\left(\hat{\theta}\right) = V\left[\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \left\{ \frac{r_{hik}}{p_c} \left(y_{hik} - \bar{Y}_c^R \right) - y_{hik} \right\} \right]$$

$$= V\left[E\left[\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \left\{ \frac{r_{hik}}{p_c} \left(y_{hik} - \bar{Y}_c^R \right) - y_{hik} \right\} \middle| \mathbf{Z} \right] \right]$$

$$+ E\left[V\left[\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \left\{ \frac{r_{hik}}{p_c} \left(y_{hik} - \bar{Y}_c^R \right) - y_{hik} \right\} \middle| \mathbf{Z} \right] \right]$$

$$= V\left(\sum_{hik\in U} Z_{hik} w_{hik} e_{Rhik} \right) + E\left[V\left[\sum_{hik\in U} Z_{hik} w_{hik} \sum_{c=1}^{C} \delta_{chik} \frac{r_{hik}}{p_c} \left(y_{hik} - \bar{Y}_c^R \right) \middle| \mathbf{Z} \right] \right].$$

Lemma 1 and Assumption (A4), which guarantees that $1/p_c$ is bounded, imply that the second term is $O\left(M^2/n\right)$.

To show that $V_2\left(\hat{\theta}\right) = o\left(M^2/n\right)$, note that by (A4) and the Cauchy-Schwarz inequality,

$$V\left[\frac{\hat{T}_c}{p_c} \right] \leq \frac{1}{\varepsilon^2} \sum_{c=1}^{C}\sum_{d=1}^{C} \sqrt{ V\left[\left(\bar{y}_c^R - \bar{Y}_c^R\right) \left(\hat{M}_c^R - M_c^R \right) \right] V\left[\left(\bar{y}_d^R - \bar{Y}_d^R\right) \left(\hat{M}_d^R - M_d^R \right) \right] }.$$

Assumption (A2) implies (Fuller 2009, Theorem 1.3.2) that

$$\sqrt{n}\begin{bmatrix} \overline{y}_c^R - \overline{Y}_c^R \\ \hat{M}_c^R / M_c^R - 1 \end{bmatrix} \to N\left(\mathbf{0}, \mathbf{\Sigma}_c\right)$$

as $n \to \infty,$ where $\mathbf{\Sigma}_c$ is a non-negative definite matrix. Consequently,

$$\left(\frac{n}{M_c^R}\right)^2 V\left[\left(\overline{y}_c^R - \overline{Y}_c^R\right)\left(\hat{M}_c^R - M_c^R\right)\right] \to \mathbf{\Sigma}_c[1,1]\,\mathbf{\Sigma}_c[2,2] + 2\left(\mathbf{\Sigma}_c[1,2]\right)^2;$$

applying the Cauchy-Schwarz inequality to the covariance term implies that $V_2\left(\hat{\theta}\right) = o\left(M^2/n\right)$.

Proof of Theorem 2. We show that

$$\tilde{V}(\theta) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in S_h} \left(\tilde{b}_{hi} - \tilde{b}_h\right)^2$$

is consistent, where

$$\tilde{b}_{hi} = \sum_{k \in S_{hi}} w_{hik} \left\{ \sum_{c=1}^{C} \frac{1}{p_c} r_{hik} \delta_{chik} \left(y_{hik} - \overline{Y}_c^R\right) - y_{hik} \right\} = \sum_{k \in S_{hi}} w_{hik} \tilde{e}_{rhik}$$

and $\tilde{b}_h = \sum_{i \in S_h} \tilde{b}_{hi} / n_h$. Arguments in Yung and Rao (2000) then imply that $(n/M^2)\left[\tilde{V}(\theta) - \hat{V}(\theta)\right]$ converges to zero in probability.

Note that

$$\begin{aligned} E\left[\tilde{b}_{hi} \mid \mathbf{Z}\right] &= \sum_{k \in S_{hi}} w_{hik} e_{Rhik}, \\ E\left[\tilde{b}_{hi}^2 \mid \mathbf{Z}\right] &= E\left[\left(\sum_{k \in S_{hi}} w_{hik} \left\{ \sum_{c=1}^{C} \frac{1}{p_c} \left[R_{hik} + r_{hik} - R_{hik}\right] \delta_{chik} \left(y_{hik} - \overline{Y}_c^R\right) - y_{hik} \right\}\right)^2 \middle| \mathbf{Z}\right] \\ &= \left(\sum_{k \in S_{hi}} w_{hik} e_{Rhik}\right)^2 + V\left(\sum_{k \in S_{hi}} w_{hik} \tilde{e}_{rhik} \middle| \mathbf{Z}\right), \end{aligned}$$

and

$$\begin{aligned} E\left[\tilde{b}_h^2 \middle| \mathbf{Z}\right] &= \frac{1}{n_h^2} E\left[\sum_{i \in S_h} b_{hi}^2 + \sum_{i \in S_h} \sum_{j \neq i} b_{hi} b_{hj} \middle| \mathbf{Z}\right] \\ &= \frac{1}{n_h^2} \sum_{i \in S_h} V\left(\sum_{k \in S_{hi}} w_{hik} \tilde{e}_{rhik} \middle| \mathbf{Z}\right) + \left(\frac{1}{n_h} \sum_{i \in S_h} \sum_{k \in S_{hi}} w_{hik} e_{Rhik}\right)^2. \end{aligned}$$

This implies that

$$\begin{aligned} E\left[\sum_{i \in S_h} \left[\tilde{b}_{hi} - \tilde{b}_h\right]^2\right] &= E\left[\sum_{i \in S_h} \left(\sum_{k \in S_{hi}} w_{hik} e_{Rhik}\right)^2 - \frac{1}{n_h} \left(\sum_{i \in S_h} \sum_{k \in S_{hi}} w_{hik} e_{Rhik}\right)^2\right] \\ &\quad + \left(1 - \frac{1}{n_h}\right) E\left[\sum_{i \in S_h} V\left(\sum_{k \in S_{hi}} w_{hik} \tilde{e}_{rhik} - y_{hik} \middle| \mathbf{Z}\right)\right], \end{aligned}$$

so that $\hat{V}_L\left(\hat{\theta}\right)$ is an approximately unbiased estimator of $V_1\left(\hat{\theta}\right)$. The consistency follows by (A2), which implies asymptotic normality, and the law of large numbers.

Proof of Theorem 3. For $c \neq d$,

$$\mathrm{Cov}\left[\left(\bar{y}_c^R - \bar{Y}_c^R\right)\left(\hat{M}_c^R - M_c^R\right),\left(\bar{y}_d^R - \bar{Y}_d^R\right)\left(\hat{M}_d^R - M_d^R\right)\right] = o\left(M^2/n^2\right)$$

because $E\left[\left(\bar{y}_c^R - \bar{Y}_c^R\right)\left(\bar{y}_d^R - \bar{Y}_d^R\right)\right] = o\left(n^{-1}\right)$ for simple random sampling (equation (4.26) of Lohr 2010). Consequently,

$$\begin{aligned} V\left(\sum_{c=1}^{C}\frac{\hat{T}_c}{p_c}\right) &= \sum_{c=1}^{C}\sum_{d=1}^{C}\frac{1}{p_c}\frac{1}{p_d}\mathrm{Cov}\left[\left(\bar{y}_c^R - \bar{Y}_c^R\right)\left(\hat{M}_c^R - M_c^R\right),\left(\bar{y}_d^R - \bar{Y}_d^R\right)\left(\hat{M}_d^R - M_d^R\right)\right] \\ &= \sum_{c=1}^{C}\frac{1}{p_c^2}V\left[\bar{y}_c^R - \bar{Y}_c^R\right]V\left[\hat{M}_c^R - M_c^R\right] + o\left(M^2/n^2\right). \end{aligned}$$

The second term of $V_2\left(\hat{\theta}\right)$ is:

$$\begin{aligned} &2\ \mathrm{Cov}\left[\sum_{c=1}^{C}\frac{\hat{T}_c}{p_c},\sum_{c=1}^{C}\frac{\left(\bar{y}_c^R - \bar{Y}_c^R\right)\hat{M}_c^R}{p_c} - \hat{Y}_{SS}\right] \\ &\quad= 2\sum_{c=1}^{C}\sum_{d=1}^{C}\frac{1}{p_c p_d}\mathrm{Cov}\left[\hat{T}_c,\left(\bar{y}_d^R - \bar{Y}_d^R\right)\hat{M}_d^R - p_d\hat{M}_d^R\bar{y}_d^R - p_d\hat{Y}_d^{NR}\right] \\ &\quad= 2\sum_{c=1}^{C}\frac{1}{p_c^2}\mathrm{Cov}\left[-\left(\bar{y}_c^R - \bar{Y}_c^R\right)\left(\hat{M}_c^R - M_c^R\right),\left(1-p_c\right)\bar{y}_c^R\hat{M}_c^R - \bar{Y}_c^R\hat{M}_c^R - p_c\hat{Y}_c^{NR}\right] + o\left(\frac{M^2}{n^2}\right) \\ &\quad= 2\sum_{c=1}^{C}\frac{p_c - 1}{p_c^2}V\left[\bar{y}_c^R - \bar{Y}_c^R\right]V\left[\hat{M}_c^R - M_c^R\right] + o\left(\frac{M^2}{n^2}\right). \end{aligned}$$

Combining the terms,

$$V_2\left(\hat{\theta}\right) = \sum_c\frac{2p_c - 1}{p_c^2}V\left[\bar{y}_c^R - \bar{Y}_c^R\right]V\left[\hat{M}_c^R - M_c^R\right] + o\left(\frac{M^2}{n^2}\right).$$

We can estimate p_c by the empirical response rate in poststratum c, $V\left[\bar{y}_c^R - \bar{Y}_c^R\right]$ by s_c^2/n_c^R, and, under simple random sampling, $V\left[\hat{M}_c^R - M_c^R\right] = M_c p_c\left(M - M_c p_c\right)/n$. The term $V_2\left(\hat{\theta}\right)$ can be negative when $p_c < 1/2$ for some poststrata; however, when $p_c < 1/2$ and $V\left[\bar{y}_c^R\right] > 0$, then the first-order term of the variance, $V_1\left(\hat{\theta}\right)$, is positive and the second-order term has lower order.

Proof of Theorem 4. Condition (A4) guarantees that, asymptotically, complete separation will not occur and R_{hik}^M is bounded away from 0.

The derivative of $\hat{\mathbf{A}}$ with respect to the parameters is

$$\begin{aligned} \hat{\mathbf{D}}\left(\mathbf{r},\boldsymbol{\beta},\theta\right) &= \frac{\partial\hat{\mathbf{A}}}{\partial\left(\boldsymbol{\beta},\theta\right)'} \\ &= \begin{bmatrix} -\sum_{hik\in S} w_{hik}\left[1+\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\right]^{-2}\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\mathbf{x}_{hik}\mathbf{x}_{hik}' & \mathbf{0} \\ -\sum_{hik\in S} w_{hik}r_{hik}y_{hik}\exp\left(-\mathbf{x}_{hik}'\boldsymbol{\beta}\right)\mathbf{x}_{hik}' & -1 \end{bmatrix}. \end{aligned}$$

Using successive conditioning and the independence of $\mathbf{r}$ and $\mathbf{Z}$, the expected value of $\hat{\mathbf{D}}(\mathbf{r},\boldsymbol{\beta},\theta)$ is

$$\mathbf{D}(\mathbf{R},\boldsymbol{\beta},\theta) = \begin{bmatrix} -\sum_{hik\in U}\left[1+\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\right]^{-2}\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\mathbf{x}_{hik}\mathbf{x}'_{hik} & \mathbf{0} \\ -\sum_{hik\in U} R_{hik}y_{hik}\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\mathbf{x}'_{hik} & -1 \end{bmatrix}$$
$$= \begin{bmatrix} -\mathbf{X}'[\mathbf{I}+\mathbf{Q}]^{-2}\mathbf{Q}\mathbf{X} & \mathbf{0} \\ -\mathbf{T}'\mathbf{Q}\mathbf{X} & -1 \end{bmatrix}.$$

Also, $\mathrm{Cov}\left[\mathrm{vec}\hat{\mathbf{D}}(\mathbf{r},\boldsymbol{\beta},\theta)\right] = O(M^2/n)$ because

$$V\left[\sum_{hik\in S} w_{hik}r_{hik}y_{hik}\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\mathbf{x}'_{hik}\right] = V\left[\sum_{hik\in S} w_{hik}R_{hik}y_{hik}\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\mathbf{x}'_{hik}\right]$$
$$+ E\left\{V\left[\sum_{hik\in S} w_{hik}r_{hik}y_{hik}\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)\mathbf{x}'_{hik}\middle|\mathbf{Z}\right]\right\}.$$

The first term is $O(M^2/n)$ by standard arguments and the second term is $O(M^2/n)$ by Lemma 1, noting that the boundedness of R_{hik} and $\mathbf{x}_{hik}$ also bound $\exp\left(-\mathbf{x}'_{hik}\boldsymbol{\beta}\right)$. Consequently,

$$V\begin{bmatrix}\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \hat{\theta}-\theta\end{bmatrix} = \mathbf{D}(\mathbf{R},\boldsymbol{\beta},\theta)^{-1}V\left[\sum_{hik\in S} w_{hik}\mathbf{u}\left(y_{hik},\mathbf{x}_{hik},r_{hik},\boldsymbol{\beta}\right)\right]\mathbf{D}(\mathbf{R},\boldsymbol{\beta},\theta)^{-T} + o(M^2/n).$$

The result in (3.3) follows because

$$[\mathbf{D}(\mathbf{R},\boldsymbol{\beta},\theta)]^{-1} = \begin{bmatrix} -\mathbf{C} & \mathbf{0} \\ \mathbf{T}'\mathbf{Q}\mathbf{X}\mathbf{C} & -1 \end{bmatrix}.$$

References

Abraham, K.G., Helms, S. and Presser, S. (2009). How social processes distort measurement: The impact of survey nonresponse on estimates of volunteer work in the United States. *American Journal of Sociology*, 114(4), 1129-1165.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on nonprobability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, 51, 279-292.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29(3), 329-353.

Eltinge, J.L. (2002). Diagnostics for the practical effects of nonresponse adjustment methods. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc., 431-443.

Fay, R.E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, DC, 429-440.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: Wiley.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 169(3), 439-457.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5), 646-675.

Hamrick, K.S. (2012). *Nonresponse Bias Analysis of Body Mass Index Data in the Eating and Health Module*. Technical Report 1934, United States Department of Agriculture Economic Research Service, Washington, DC.

Harris-Kojetin, B.A. (2012). *Nonresponse Bias Analysis for Establishment Surveys: Guidance from the U.S. Office of Management and Budget*. Paper presented at DC-AAPOR, http://www.dc-aapor.org/documents/Harris-Kojetin2012.pdf.

Haziza, D., and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32(1), 129-145.

Haziza, D., Thompson, K.J. and Yung, W. (2010). The effect of nonresponse adjustments on variance estimation. *Survey Methodology*, 36, 1, 35-43. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2010001/article/11246-eng.pdf.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics*, 35(4), 501-514.

Kohut, A., Keeter, S., Doherty, C., Dimock, M. and Christian, L. (2012). *Assessing the Representativeness of Public Opinion Surveys*. Pew Research Center, Washington, DC.

Lohr, S.L., Hsu, V. and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 2071-2085.

Oh, H.L., and Scheuren, F.J. (1987). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys,* (Eds., W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 2, 143-184.

Riddles, M., Marker, D.A., Rizzo, L., Wiley, E. and Zukerberg, A. (2015). *Adaptive Design for the National Teacher Principal Survey*. Paper presented at the AAPOR 70th Annual Conference, Hollywood, FL.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Hoboken, NJ: Wiley.

SAS Institute, Inc. (2011). *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute, Inc.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 93, 254-265.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2016). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, published online, http://onlinelibrary.wiley.com/doi/10.1111/rssa.12186/pdf.

United States Office of Management and Budget (2006). Standards and guidelines for statistical surveys. https://www.whitehouse.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Yung, W., and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451), 903-915.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 219-238
Statistics Canada, Catalogue No. 12-001-X

Reducing the response imbalance: Is the accuracy of the survey estimates improved?

Carl-Erik Särndal, Kaur Lumiste and Imbi Traat[1]

Abstract

We present theoretical evidence that efforts during data collection to balance the survey response with respect to selected auxiliary variables will improve the chances for low nonresponse bias in the estimates that are ultimately produced by calibrated weighting. One of our results shows that the variance of the bias – measured here as the deviation of the calibration estimator from the (unrealized) full-sample unbiased estimator – decreases linearly as a function of the response imbalance that we assume measured and controlled continuously over the data collection period. An attractive prospect is thus a lower risk of bias if one can manage the data collection to get low imbalance. The theoretical results are validated in a simulation study with real data from an Estonian household survey.

Key Words: Survey nonresponse; Bias; Adaptive data collection; Calibration estimator; Auxiliary variables.

1 Introduction

The problem of accurate estimation despite considerable nonresponse needs to be examined from two time dependent angles: First, ways to handle the data collection, then ways to handle the estimation with the data that were finally collected. The first activity may require substantial resources. In a telephone survey, the daily scheduling of contact attempts, the interaction with the interviewers, and consideration for their workloads, can be expensive efforts. The estimation stage is administratively simpler; there is a search for the best auxiliary variables for a calibrated nonresponse adjustment weighting, whereupon the computation of estimates is usually carried out with existing software.

The data collection is in focus in the literature on Responsive Design; Groves (2006), Groves and Heeringa (2006) are early references. Adaptive survey designs are discussed in Wagner (2008). One idea in this research tradition is that a data collection that extends over a period of time might be inspected at suitable decision points, where action may be taken to realize in the end a well-balanced set of respondents. Schouten, Calinescu and Luiten (2013) explain how adaptive survey designs may be tailored to optimize response rates and reduce nonresponse selectivity, with cost aspects taken into account. Much exploratory work has been carried out on responsive (or adaptive) design. Seeking well balanced or representative response can be pursued as a goal in itself. Different avenues have been explored: Case prioritization, (Peytchev, Riley, Rosen, Murphy and Lindblad 2010); stopping rules to halt data collection attempts for specific sample units, (Rao, Glickman and Glynn 2008; Wagner and Raghunathan 2010); uses of paradata more generally to manage the survey response, (Couper and Wagner 2011).

Measuring and controlling the imbalance belongs in the data collection phase. The imbalance statistic (see Section 3) has a central role in this article; it was used for example in Särndal (2011), Lundquist and Särndal (2013), Särndal and Lundquist (2014a, 2014b). It is related to the R- indicator (R for

1. Carl-Erik Särndal, Ph.D., professor emeritus, Statistics Sweden. E-mail: carl.sarndal@telia.com; Kaur Lumiste, M.Sc., Institute of Mathematical Statistics, University of Tartu, Estonia; Imbi Traat, Ph.D., associate professor, Institute of Mathematical Statistics, University of Tartu, Estonia.

representativity); see Schouten, Cobben and Bethlehem (2009) and Bethlehem, Cobben and Schouten (2011).

The second time slice relies on estimation theory to resolve the challenge of nonresponse, primarily how to achieve low bias in the estimates. Viewed strictly as an estimation problem, it is an activity in itself, after a completed data collection. The set of responding units is fixed; the data on those units is a "frozen" supply. The choice of auxiliary variables plays a crucial role. The "best ones" should be selected. This aspect has been dealt with extensively, as in Särndal and Lundström (2005). Two factors are traditionally cited as important for the accuracy of the estimates: The degree to which the chosen auxiliary variables can explain the study variable and the degree to which these variables can explain the 0/1 response indicator showing presence or not in the set of respondents. Each of the two degrees of explanation is partial at best, not perfect. The two roles of the auxiliary variables interact, as recognized for example in Little and Vartivarian (2005). An extensive review of weighting adjustment procedures for nonresponse is given in Brick (2013).

The supply of auxiliary variables depends on the survey environment. In Scandinavia, surveys on individuals and households can draw on extensive sources – administrative registers – of auxiliary variables. This is increasingly so in other countries also.

One view holds that the estimation is the all-important step: Whatever may be accomplished at the data collection stage – balancing, improved representativeness – is perhaps superfluous; achieving best possible accuracy in the estimates can be dealt with effectively at the estimation stage, by clever use of the auxiliary variables in a nonresponse adjustment weighting or in other ways. This point of view is supported for example in Beaumont, Bocci and Haziza (2014).

Nevertheless, it is clear that measurable aspects of the data collection will influence the accuracy of the estimates that are ultimately produced. One such measure is the imbalance statistic defined in Section 3. In this article, the two time dependent activities are taken into account: Balancing the response should be combined with efficient estimation methods, to get in the end the best possible (most accurate) estimates. Such a view underlies, for example, Schouten, Cobben, Lundquist and Wagner (2014).

The motivation for this paper is as follows: Methods exist for different courses of action – stopping rules, case prioritization, and others – during data collection, so as to get in the end a favourable response set r. Särndal and Lundquist (2014a, 2014b) used the imbalance statistic IMB given in Section 3 as a tool to achieve low imbalance in the final response set. Considering that auxiliary variables will also be used in the estimation, to what extent, if any, will better accuracy in the estimates follow from low imbalance in the preceding data collection? There are encouraging signs, as in Särndal and Lundquist (2014a), that lower imbalance creates some accuracy improvement, although modest. That work was empirical; in this article we give mathematical/analytical support for a similar conclusion.

The contents are arranged as follows: The survey background (Section 2) and the imbalance statistic (Section 3) are presented. The regression relationship – that of the study variable on the auxiliary vector – is important (Section 4), notably for the estimator (called CAL) obtained by calibrated nonresponse weight adjustment (Section 5). The deviation of the calibration (CAL) estimator from the (unbiased) estimator requiring full response is analyzed (Sections 6, Section 7, Section 8), showing how deviation depends on imbalance. Two results are presented on statistical properties (mean and variance) of the CAL deviation. In

particular, the variance of that deviation is shown to be, approximately, a linear function of the imbalance statistic. Hence the deviation is likely to be smaller, and estimates more accurate, if the imbalance can be reduced during data collection. The theoretical results are empirically validated (Section 9) using data from an Estonian household survey. The statistical software R is used; R Core Team (2014). A discussion (Section 10) concludes the article. Three appendices provide the necessary proofs and derivations.

2 Background and notation

A probability sample s is drawn from the finite population $U = \{1, 2, \ldots, k, \ldots, N\}$. Unit k has the known inclusion probability $\pi_k = \Pr(k \in s)$ and the known design weight $d_k = 1/\pi_k$. Nonresponse occurs. The response set, denoted r, is that subset of s for which the study variable is observed. We do not know how r was generated from s; the response probabilities are unknown (if assumed to "exist", they are not needed in this article). The (design weighted) response rate is

$$P = \sum_r d_k \Big/ \sum_s d_k . \tag{2.1}$$

If A is a set of units, $A \subseteq U$, a sum $\sum_{k \in A}$ will be written as $\sum_A$. The survey may have many study variables. A typical one, denoted y (continuous or categorical), has value y_k recorded for $k \in r$ but missing for $k \in s - r$. Our objective is to estimate the population y-total, $Y = \sum_U y_k$. The response indicator I has value $I_k = 1$ for $k \in r$, $I_k = 0$ for $k \in s - r$. A goal for practice is to get a response r that is well balanced, in the sense specified later. We are led to consider the different r that may arise from a given s.

The auxiliary vector $\mathbf{x}$ of dimension $J \geq 1$ has value $\mathbf{x}_k$ known at least for all units $k \in s$. Auxiliary information can be used in the data collection (for monitoring the data inflow to achieve improved balance) and/or in the estimation (for calibrated weight computation). The auxiliary vector need not be the same for the two purposes, but this article assumes that they agree, and that the $\mathbf{x}$-information used is for $k \in s$. This includes the important case of paradata, that is, data about the data collection process.

An important type of auxiliary vector is a *group vector*. It identifies membership of every unit k in one of J mutually exclusive and exhaustive sample groups, so that $\mathbf{x}_k = (0, \ldots, 1, \ldots 0)'$, where the only "1" indicates the unique group (out of J possible) to which k belongs.

A group vector occurs when several categorical auxiliary variables are completely crossed. To illustrate, if $\mathbf{x} = (sex \times education \times age)$ represents a crossing of 2 sexes, 3 exhaustive education categories and 4 exhaustive age categories, then $\mathbf{x}$ is a group vector with dimension $J = 2 \times 3 \times 4 = 24$ and equally many possible values $\mathbf{x}_k$. When several categorical variables are used although not in completely crossed manner – important for practice in statistical agencies – then the dimension J of $\mathbf{x}_k$ can be kept relatively modest (say less than 15) while still coding a much larger number (say more than one hundred) of possible properties $\mathbf{x}_k$ of the units k. For a study of the Swedish Living Conditions Survey, Särndal and Lundquist

(2014a) used an $\mathbf{x}$-vector of dimension 14 with 256 possible values. The group vector case and the non-group vector case give important differences in the results that follow.

All auxiliary vectors used here satisfy a requirement that grants mathematical convenience without severely restricting the choice of vector: There exists a constant vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k. \tag{2.2}$$

For example, when $J = 2$ and $\mathbf{x}_k = (1, x_k)'$, then $\boldsymbol{\mu} = (1,0)'$ satisfies the requirement. In the group vector case where $\mathbf{x}_k = (0,\ldots,1,\ldots,0)'$, then $\boldsymbol{\mu} = (1,\ldots,1,\ldots,1)'$ satisfies the requirement. If $\mathbf{x}$ is not a group vector, say, one used to code "education" with three mutually exclusive and exhaustive categories and "gender" as a univariate variable equal to 1 or 0, then $J = 3+1 = 4$ (education and gender not crossed), and $\boldsymbol{\mu} = (1,1,1,0)'$ satisfies the requirement.

3 Imbalance

The concept of *balance* has been often used in statistical literature with reference to an equality of means of specified variables for two sets of units, one a subset of the other. One method to realize a probability sample s from U that is balanced with respect to a vector $\mathbf{x}$ is the Cube Method, see Deville and Tillé (2004). In the context with nonresponse, we want to know how well balanced a response r is, compared with the probability sample s that would have given unbiased estimates. A given auxiliary vector $\mathbf{x}$ has computable means $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k \big/ \sum_r d_k$ for the response and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k \big/ \sum_s d_k$ for the sample. If they are equal, an unlikely outcome, the response is perfectly balanced with respect to $\mathbf{x}$. The contrast between response r and sample s can be measured by the scalar quantities

$$Q_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right); \quad Q_r = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_r^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right). \tag{3.1}$$

They differ only in the $J \times J$ weighting matrix, $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' \big/ \sum_s d_k$ as opposed to $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' \big/ \sum_r d_k$, both assumed non-singular. In particular, Q_s is important for the statistic called *imbalance* of r with respect to the specified $\mathbf{x}$-vector:

$$IMB(r, \mathbf{x} \mid s) = P^2 \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \boldsymbol{\Sigma}_s^{-1} \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right) = P^2 Q_s, \tag{3.2}$$

where P is the response rate (2.1); see for example Särndal and Lundquist (2014a). The full notation $IMB(r, \mathbf{x} \mid s)$ emphasizes that imbalance depends on the realized response r and on the choice of $\mathbf{x}$-vector. Unless required for emphasis, we use the simpler notation IMB. We have $0 \le IMB \le P(1-P)$ for any r and vector formulation $\mathbf{x}$, given s. IMB is a descriptive measure of the response r. It is related to a special case of the R-indicator, whose motivation lies instead in the estimation of (the unknown) response probabilities for the population units, see for example Bethlehem et al. (2011).

The IMB statistic (3.2) can be continuously computed and monitored in a data collection extending over a period of time, say several days or weeks, during which contact attempts continue with a sample unit until desired data are obtained, or, if this fails, until the unit is declared a non-respondent. As the response rate P grows, IMB serves as a tool for monitoring and managing the data collection to achieve in the end a response set r which, if not perfectly balanced to satisfy $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, will at least have considerably lower IMB than if balancing had not been attempted in data collection. There are methods for balancing based on response propensity, such as the Threshold method and the Equal proportions method in Särndal and Lundquist (2014a, 2014b).

We consider later the particular case where s is a self-weighting sample (as when s is a simple random sample), the response r has fixed size m, and $\mathbf{x}$ is a group vector of dimension J as defined in Section 2. Then both s and r are split into J non-overlapping groups. For the sample group s_j, denote by n_j the size and by $W_{js} = n_j / n$ the relative size; $\sum_{j=1}^{J} n_j = n$. For the response group r_j, let $m_j \le n_j$ be the size; $\sum_{j=1}^{J} m_j = m$. The imbalance (3.2) is then

$$IMB = \sum_{j=1}^{J} W_{js} \left(p_j - p \right)^2, \tag{3.3}$$

where the response rates are $p_j = m_j / n_j$ in group j and $p = m/n$ overall. (The response rate P is defined in (2.1) with general design weights d_k; for a self-weighting sample, where d_k is constant, we use small p for the response rate.) If $IMB = 0$, we have perfect balance; all group response rates p_j are then equal.

4 The regression aspect

The imbalance (*IMB*) is determined by the auxiliary vector $\mathbf{x}$ with no attention paid to the study variable y. But the relation of $\mathbf{x}$ to y is also important for the bias of estimated y-totals. Strong regression of y on $\mathbf{x}$ is likely to give small bias, intuitively because regression predicted y-values can then give close substitutes for those missing. For some survey data, the strength of the regression may be modest but nevertheless important in its effect on bias. The ordinary linear regression coefficient vectors for the whole sample s and for the response r are, respectively,

$$\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_s d_k \mathbf{x}_k y_k; \quad \mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_r d_k \mathbf{x}_k y_k. \tag{4.1}$$

Under nonresponse, $\mathbf{b}_r$ is computable but not $\mathbf{b}_s$. The $J \times J$ matrices to invert are assumed non-singular. Normally $\mathbf{b}_r \ne \mathbf{b}_s$, perhaps with considerable (but unknown) difference. The regression based on the response is inconsistent.

The imbalance in the y-variable is $\bar{y}_r - \bar{y}_s$, where the means are $\bar{y}_s = \sum_s d_k y_k / \sum_s d_k$ for the sample (unknown) and $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$ for the response (computable). The decomposition

$$\bar{y}_r - \bar{y}_s = \left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r + \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s, \tag{4.2}$$

highlights two undesirable differences, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ (due to imbalance in the $\mathbf{x}$-vector), and $\mathbf{b}_r - \mathbf{b}_s$ (due to inconsistent regression); to obtain (4.2) note that $\bar{\mathbf{x}}_r'\mathbf{b}_r = \bar{y}_r$ and $\bar{\mathbf{x}}_s'\mathbf{b}_s = \bar{y}_s$, which are consequences of the $\mathbf{x}$-vector condition (2.2).

5 Estimating the population total under nonresponse

The equation (4.2), when multiplied by $\hat{N} = \sum_s d_k$, can be expressed in terms of three common estimators of the population total $Y = \sum_U y_k$. Two are possible under nonresponse,

$$\hat{Y}_{EXP} = \hat{N} \frac{\sum_r d_k y_k}{\sum_r d_k}, \quad \hat{Y}_{CAL} = \hat{N} \frac{\sum_r d_k g_k y_k}{\sum_r d_k}, \tag{5.1}$$

with $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. Of these, $\hat{Y}_{EXP}$ is just a simple expansion of the response mean of y and often considerably biased. The calibration estimator $\hat{Y}_{CAL}$ gives y_k the weight $d_k g_k / P$. The calibration property is $\sum_r \left(d_k g_k / P\right) \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, where the right hand side is unbiased for the population $\mathbf{x}$-total $\sum_U \mathbf{x}_k$, which explains why $\hat{Y}_{CAL}$ can be considerably less biased than $\hat{Y}_{EXP}$ when $\mathbf{x}$ and y are well related. If y-values had been recorded for the full sample s, unbiased estimation would be carried out with the Horvitz-Thompson estimator

$$\hat{Y}_{FUL} = \sum_s d_k y_k.$$

The three estimator types will be referred to as EXP, CAL and FUL. Now (4.2) multiplied by $\hat{N} = \sum_s d_k$ reads

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = \left(\hat{Y}_{EXP} - \hat{Y}_{CAL}\right) + \left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right). \tag{5.2}$$

In words, Deviation of $EXP =$ Bias adjustment term + Deviation of CAL. The computable adjustment is $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$. The two deviations from the unbiased estimate, $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ for CAL and $\hat{Y}_{EXP} - \hat{Y}_{FUL} = \hat{N}\left(\bar{y}_r - \bar{y}_s\right)$ for EXP, are not computable under nonresponse, because they require y-values for the full sample.

As mentioned, we have methods to reduce the imbalance IMB during data collection. Low imbalance is intuitively attractive, but does it yield better accuracy in estimates? Or is it enough to involve the auxiliary variables at the estimation stage, through a calibrated weight adjustment as in the CAL estimator? The adjustment term $\hat{Y}_{EXP} - \hat{Y}_{CAL} = \hat{N}\left(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s\right)' \mathbf{b}_r$ can clearly be reduced by constructing r to have low imbalance; it is zero for the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. In practice, the CAL estimator is preferred to the EXP estimator, the former being usually more accurate because of the auxiliary information. But is the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s$ smaller if the response r had been built to have low IMB? Asked

differently, is it worth the (perhaps costly) effort to manage the data collection to get $\bar{\mathbf{x}}_r$ closer to $\bar{\mathbf{x}}_s$ and therefore reduced IMB? The question is essentially whether this would also make $\mathbf{b}_r$ and $\mathbf{b}_s$ move closer.

6 Statistical properties of the CAL estimator deviation

In the decomposition (5.2), the deviation of CAL from the unbiased FUL is $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ where $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s$. To see if $\boldsymbol{\Delta}_r$ is smaller, or likely to be so, by realizing low imbalance in data collection, we seek analytic results about statistical properties, such as mean and variance, of $\boldsymbol{\Delta}_r$ as a function of the IMB statistic (3.2). Highly general results of this kind are hard to obtain. Several factors complicate the analysis, such as the sampling design used to draw s, the probability distribution of the response sets r given s, the make-up of the auxiliary vector $\mathbf{x}$, and so on. Results for special situations are obtained in Sections 7 and 8.

Result 1 in Section 7 gives properties – expected value and variance – of $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)' \bar{\mathbf{x}}_s$ over response outcomes r with fixed size m and fixed mean $\bar{\mathbf{x}}_r$ when $\mathbf{x}$ is a group vector, and s is a simple random sample. The mean of $\boldsymbol{\Delta}_r$ over such outcomes is zero. The imbalance appears in the variance of $\boldsymbol{\Delta}_r$, which is linearly increasing in IMB, approximately. A reason for taking $\mathbf{x}$ to be a group vector is that conditioning on $\bar{\mathbf{x}}_r$ grants relatively simple derivations. A fixed $\bar{\mathbf{x}}_r$ implies a fixed value IMB. (But the opposite is not true; several $\bar{\mathbf{x}}_r$ can give the same IMB.) Another simplification when $\mathbf{x}$ is a group vector is due to diagonal matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$. The empirical test in Section 9.1 addresses Result 1.

Simple derivations for the group vector are at the expense of generality. The $\mathbf{x}$-vectors used in production at Statistics Sweden, for example, are often not group vectors. To get transparent mathematical results about $\boldsymbol{\Delta}_r$ is then more difficult.

Result 2 in Section 8 is derived under a model of linear regression between y and $\mathbf{x}$. The y_k are then considered random, with properties stated by the model. A group vector feature for $\mathbf{x}$ is no longer necessary. The conclusions are in some respects similar to those in Result 1. The empirical Test situation 2 in Section 9.2 refers to both Results 1 and 2.

7 The first result

Result 1 refers to the following survey context: A self-weighting sample s of size n is drawn from $U = \{1,\ldots,k,\ldots,N\}$; d_k is the same for all k. The auxiliary vector $\mathbf{x}$ is a group vector of dimension J, so the sample s and the response set r, assumed to be of fixed size $m < n$, are split into J non-overlapping groups. The notation for these is given at the end of Section 3. The values y_k are treated as fixed, non-random, as is usual in the design-based tradition. If y_k were observed for all $k \in s$, then $\hat{Y}_{FUL} = N\,\bar{y}_s$ with $\bar{y}_s = \sum_s y_k \big/ n$ would be design unbiased for the population y-total $Y = \sum_U y_k$. But y_k is available for

$k \in r$ only; the CAL estimator (5.1) becomes $\hat{Y}_{CAL} = N\sum_{j=1}^{J} W_{js}\bar{y}_{r_j}$, where $\bar{y}_{r_j}$ is the mean of respondent values y_k in group j. Statistical properties – expected value and variance – of $\left(\hat{Y}_{CAL} - \hat{Y}_{FUL}\right)/N = \Delta_r$ with $\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js}\bar{y}_{r_j} - \bar{y}_s$ are given in Result 1 for the following probabilistic setting: All $\binom{n}{m}$ response sets r of fixed size m are assumed a priori equally probable. Given s, the imbalance IMB is determined by $\bar{\mathbf{x}}_r = (1/m)\left(m_1,\ldots,m_j,\ldots,m_J\right)'$. Given $\bar{\mathbf{x}}_r$, we are left with $R = \prod_{j=1}^{J}\binom{n_j}{m_j}$ sets r, all with the same non-zero probability $1/R$ and the same IMB, given by (3.3). The other sets r of size m are no longer in scope. Conditioning on $\bar{\mathbf{x}}_r$ allows us to study the properties of CAL as a function of IMB. Result 1 involves the variance of the study variable y, within-group and combined over groups:

$$S_{yj}^2 = \sum\nolimits_{s_j}\left(y_k - \bar{y}_{s_j}\right)^2 \Big/ \left(n_j - 1\right); \quad S_y^2 = \sum_{j=1}^{J} W_{js}S_{yj}^2. \tag{7.1}$$

Result 1. Let s be a self-weighting sample of size n and let $\mathbf{x}_k$ be a group vector of dimension J. Assume that all $\binom{n}{m}$ response sets r of fixed size m are a priori equally probable. Then

$$\bar{\Delta} = \mathrm{E}\left(\Delta_r \,|\, \bar{\mathbf{x}}_r, m, s\right) = 0 \tag{7.2}$$

$$S_\Delta^2 = \mathrm{E}\left(\left(\Delta_r - \bar{\Delta}\right)^2 \,\middle|\, \bar{\mathbf{x}}_r, m, s\right) = \left(\frac{1}{m} - \frac{1}{n}\right)S_y^2 + \frac{1}{m}\sum_{j=1}^{J} W_{js}\left(\frac{p}{p_j} - 1\right)S_{yj}^2 \tag{7.3}$$

where $W_{js} = n_j/n$ and $p_j = m_j/n_j$ are relative size and response rate, respectively, for group j, $p = m/n$ is the overall response rate, and S_y^2 and S_{yj}^2 are given in (7.1). If response rates p_j and variances S_{yj}^2 vary by little only over the groups, then

$$S_\Delta^2 \approx \left(1 - p + \frac{IMB}{p^2}\right)\frac{S_y^2}{m} \tag{7.4}$$

where IMB is given by (3.3).

For full response, when $r = s$, the right hand sides of (7.3) and (7.4) are zero; the approximation in (7.4) is exact: $S_\Delta^2 = 0$. To interpret Result 1, note that the first term on the right hand side of (7.3) is a constant, given m. It states the conditional variance for a perfectly balanced response, where p_j is the same for all groups. The second is the *penalty term*, namely the penalty for failing to get perfect balance in data collection. Its size depends on how well an adaptive design succeeds in generating group response rates p_j that vary little only. It is zero if all p_j can be made equal.

Formula (7.4) states that the variance S_Δ^2 is decreasing with IMB in a roughly linear fashion. Thus low imbalance brings improved chances for a small deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$. This is important for practice. To illustrate, for a nonresponse of $1 - p = 40$ per cent, $S_\Delta^2 \approx 0.57S_y^2/m$ if $IMB = 0.06$, but if $IMB = 0$, as in perfect balance, that variance is reduced to $S_\Delta^2 \approx 0.40S_y^2/m$. The improvement is clear but cannot be claimed to be very large. This is because with most data, IMB/p^2 is small compared with a nonresponse $1 - p$ of the order of 30 to 60 per cent, cases that we are mainly concerned with here. Thus taking action to reduce imbalance has a desirable effect, although modest rather than strong.

In (7.2) and (7.3), the expectation $\mathrm{E}(\cdot)$ is taken by averaging over the $R = \prod_{j=1}^{J} \binom{n_j}{m_j}$ equi-probable sets r that remain out of $\binom{n}{m}$ after fixing $\bar{\mathbf{x}}_r$. It should also be noted that more than one $\bar{\mathbf{x}}_r$ can give the same value IMB. Hence there may be more than one value S_Δ^2 for the same IMB. The linearly increasing function of IMB in (7.4) is nevertheless their common approximation.

8 The second result

In Result 1, the survey variable values y_k are treated as fixed, nonrandom. In Result 2, they are random with properties as stated in a linear regression model ξ with residuals $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ for some unknown $\boldsymbol{\beta}$:

$$E_\xi(y_k \mid \mathbf{x}_k) = \mathbf{x}_k'\boldsymbol{\beta}; \quad E_\xi(\varepsilon_k^2 \mid \mathbf{x}_k) = \sigma_\varepsilon^2, \text{ all } k \in s; \quad E_\xi(\varepsilon_k \varepsilon_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell) = 0, \text{ all } k \neq \ell \in s. \tag{8.1}$$

The properties in (8.1) apply also to units k and ℓ belonging in any subset r of s. Result 2 presents expected value and approximate variance of $\boldsymbol{\Delta}_r = (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ conditionally on a fixed self-weighting sample s and a fixed response set r with respective sizes n and m.

Result 2: Let s of size n be a self-weighting sample. Let $\mathbf{X}$ be the $J \times n$ $\mathbf{x}$-data matrix with columns $\mathbf{x}_k$, $k \in s$. Then, under the model ξ in (8.1),

$$E_\xi(\Delta_r \mid \mathbf{X}, r, s) = 0; \quad E_\xi(\Delta_r^2 \mid \mathbf{X}, r, s) \approx \left(1 - p + \frac{IMB}{p^2}\right)\frac{\sigma_\varepsilon^2}{m}, \tag{8.2}$$

where m is the size of the fixed response set r, $p = m/n$ is the response rate and IMB is given by (3.2).

Result 2 (for arbitrary $\mathbf{x}$-vector and random y_k) mirrors Result 1 (for group $\mathbf{x}$-vector and non-random y_k) in that both give conditional mean zero and the same linearly increasing form for the conditional variance approximation.

The derivation in Appendix 3 of Result 2 relies on a comparison of the two quadratic forms in $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ given in (3.1), Q_s and Q_r. The former is used in the imbalance statistic (3.2), $IMB = P^2 Q_s$; the latter determines the weight factors g_k for the CAL estimator (5.1). The approximation $Q_r \approx Q_s$, needed for Result 2, is justified in Appendix 2.

9 Empirical testing

Results 1 and 2 give the basis for testing empirically in this section how mean and variance of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r = \hat{N}(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ depend on the imbalance IMB. Both results state that the variance of $\boldsymbol{\Delta}_r$ increases in a roughly linear fashion as IMB increases, without being small even if IMB is near zero.

We use real data from an Estonian survey with 17,540 households. The following variables are available for every household: Household net income, used here as the study variable y, and three categorical variables referring to the designated head of household, used here as auxiliary variables: (i) Gender (1 for male, 0 for female), (ii) Economic activity (1 for employed, 0 for not employed) and (iii) Education, with three exhaustive levels: low, medium, high.

We compute the mean $\bar{\Delta}$ of $\mathbf{\Delta}_r$ and the variance S_Δ^2 of $\mathbf{\Delta}_r$ by averaging over the sets r with fixed mean $\bar{\mathbf{x}}_r$, given s.

9.1 Test situation 1

In line with Result 1, we want to consider the response sets r with fixed size m arising from a given sample s of size n. The computational volume is prohibitive even for rather small n. We drew s as a simple random sample of size $n = 20$ from 17,540. The d_k are then constant. The sample mean for the y-variable (household income) was $\bar{y}_s = 10,386.65$. We define $\mathbf{x}_k$ as the group vector of dimension $J = 3$ that identifies the three exhaustive levels of Education; low, medium, high. For the realized sample s, we have $n\bar{\mathbf{x}}_s = (5,8,7)'$.

We fixed the size of the response sets r to be $m = 12$. The response rate is 60 per cent for every one of the $\binom{20}{12} \approx 1.26 \times 10^5$ possible response sets r. From these, we excluded all those for which the response count vector $m\bar{\mathbf{x}}_r$ contained a zero, to avoid a singular $\mathbf{\Sigma}_r$. This left 31 configurations (m_1, m_2, m_3) such that $m_1 + m_2 + m_3 = 12$ and all three counts $m_j \geq 1$. For each of the 31 possibilities, we computed $\bar{\Delta}$ and S_Δ^2 by averaging over the response sets r satisfying the fixed configuration. For example, $(m_1, m_2, m_3) = (3,4,5)$ is satisfied by 14,700 response sets r, so mean and variance of Δ_r are computed over those. Other configurations give much fewer response sets, for example, only 70 for the configuration (3, 8, 1); a few of those can then be very influential in the computations. For every one of the 31 cases, $\bar{\Delta}$ is theoretically zero, by Result 1. The computations confirmed this; a plot of $\bar{\Delta}$ against IMB is unnecessary. Figure 9.1 shows the 31 point plot of S_Δ^2 against IMB. Because of the non-uniqueness of IMB noted earlier, it happens several times that more than one S_Δ^2 occurs at the same IMB value. Figure 9.1 shows that S_Δ^2 has a clear upward trend as IMB increases. Figure 9.1 also shows the approximation $S_\Delta^2 \approx S_{\Delta approx}^2 = \left(S_y^2 / m\right)(1 - p + IMB / p^2)$ from Result 1. We have $p = 0.6, m = 12$ and $S_y^2 = 26.3 \times 10^6$, so the computed approximation, linear in IMB, is $S_{\Delta approx}^2 = a + b\, IMB$ with $a = 0.879 \times 10^6$ and $b = 6.102 \times 10^6$. For points with low IMB, S_Δ^2 agrees closely with the linearly increasing $S_{\Delta approx}^2$. A contributing reason is that when IMB is low, the group response rates p_j vary little, and this is one of the conditions for close approximation, as the derivation of Result 1 in Appendix 1 explains. For higher IMB values, the increasing trend in S_Δ^2 is still evident, but the scatter around the theoretical line is more pronounced. Five outlying points in Figure 9.1 have very large S_Δ^2; three of them occur when one component of (m_1, m_2, m_3) is equal to the maximal count (5 or 8 or 7). For those, less accurate linear approximation is expected, the p_j being far from equal.

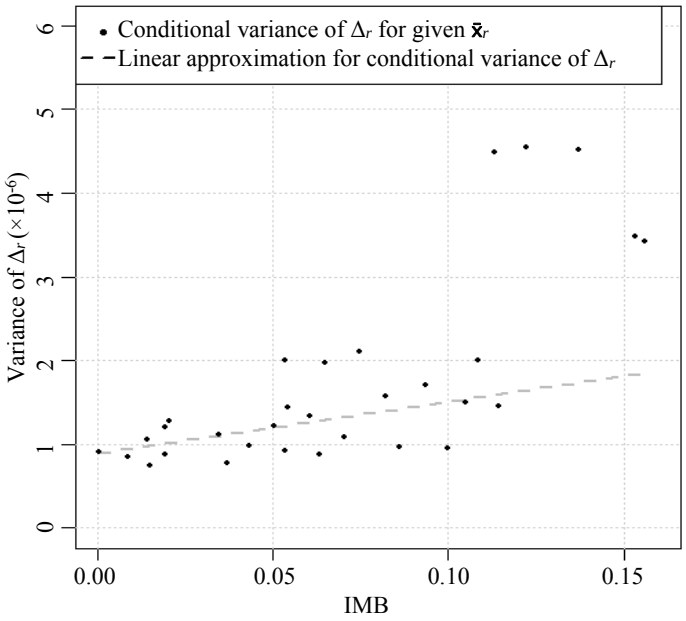

Figure 9.1 Conditional variance of Δ_r as a function of imbalance *IMB*; $\mathbf{x}_k$ a group vector of dimension 3; response sets r of fixed size 12 from a fixed sample s of size 20.

9.2 Test situation 2

The setup and the computational steps are similar to those in Test situation 1, but $\mathbf{x}_k$ is no longer a group vector; some results change considerably, compared with Test situation 1.

A new simple random sample s of size $n = 20$ was drawn from the 17,540 households. For this sample, $\bar{y}_s = 9{,}618.4$. We let $\mathbf{x}_k$ incorporate all three auxiliary variables (i), (ii) and (iii), but not completely crossed: Gender (univariate coded 0 or 1), Economic activity (univariate coded 0 or 1) and Education level (three exhaustive categories coded (1,0,0) or (0,1,0) or (0,0,1)). This $\mathbf{x}_k$ is not a group vector; it has dimension $1+1+3=5$ and $2\times 2\times 3 = 12$ possible values; $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$ are not diagonal. We have $n\bar{\mathbf{x}}_s = (9,11,4,7,9)'$. For this sample s we considered the response sets r of fixed size $m = 12$ excepting those where one or more of the five components of the count vector $m\bar{\mathbf{x}}_r$ are zero. This left 658 different vectors $m\bar{\mathbf{x}}_r$, each composed of five non-zero counts, and satisfied by a certain number of response sets r over which we computed, by simple averaging, the mean $\bar{\Delta}$ and variance S^2_{Δ}. These are thus moments conditionally on $\bar{\mathbf{x}}_r$.

Figure 9.2 shows the 658 point plot of $\bar{\Delta}$ against *IMB*. In Test situation 1, $\bar{\Delta}$ was zero for every point because $\mathbf{x}_k$ was a group vector. This is not so in Figure 9.2, where the means $\bar{\Delta}$ fan out when *IMB* increases. They are much more concentrated around zero for low *IMB* than for large *IMB*. Several points (that is several means $\bar{\mathbf{x}}_r$) can give the same or nearly the same *IMB*. Figure 9.2 shows that in a small neighborhood of a fixed value IMB_0 on the *IMB* axis, the mean of the means $\bar{\Delta}$ is roughly zero. With reference to Result 2, we can expect to see the average of $\bar{\Delta}$ for fixed *IMB* to be near zero: Under model

(8.1) for y_k, Result 2 says that $E_\xi(\Delta_r \mid \mathbf{X}, r, s) = 0$. When $\mathbf{X}$ and r are fixed, so is IMB. If the model is a reasonably good representation, the average of Δ_r for fixed IMB should be close to zero, as Figure 9.2 indicates.

Figure 9.3 shows the plot of the conditional variance S^2_Δ against IMB. The pattern with a variance S^2_Δ that increases linearly in IMB prevails, even though $\mathbf{x}_k$ is not a group vector here. Figure 9.3 shows the computed approximating line $S^2_{\Delta approx} = (\hat{\sigma}^2_\varepsilon / m)(1 - p + IMB/p^2)$ derived from Result 2, with $\hat{\sigma}^2_\varepsilon = \sum_s (y_k - \mathbf{x}'_k \mathbf{b}_s)^2 / (n - J)$ used to estimate σ^2_ε. We have $J = 5$, $p = 0.6$, $m = 12$ and $\hat{\sigma}^2_\varepsilon = 33.6 \times 10^6$, so the line in Figure 9.3 is $S^2_{\Delta approx} = a + b\, IMB$ with $a = 1.12 \times 10^6$ and $b = 7.78 \times 10^6$. The linear approximation holds particularly well for small IMB, say less than 0.1. For large IMB, there is much scatter; S^2_Δ has some very large values, and some very low values as well. Figure 9.4 shows the joint behavior of $\bar{\Delta}$ and S^2_Δ for the 658 points. The size of a dot is proportional to IMB^2; the reason for squaring is to better contrast larger and smaller IMB values. Response sets r with small IMB are found to give small $\bar{\Delta}$ and S^2_Δ, a favourable sign because the CAL and FUL estimators are then close. To illustrate, for points satisfying $IMB \le 0.1$, $\bar{\Delta}$ is in the interval (-1,390; 1,447) and S^2_Δ in (0.846×10^6; 4.86×10^6). These are narrow intervals; this is even more pronounced for $IMB \le 0.05$. But when IMB is large, this advantageous situation no longer holds. For example, $\bar{\Delta}$ can be very small and at the same time S^2_Δ very large (points in the middle and right side of the figure). On the other hand, S^2_Δ can be near zero while $\bar{\Delta}$ is very large in absolute value (points in the top and bottom left parts of the figure.) Test situation 2 illustrates that a non-group vector $\mathbf{x}_k$ can give both a distinctly non-zero mean of Δ_r and a high variance of Δ_r, and that these tendencies are accentuated by large imbalance.

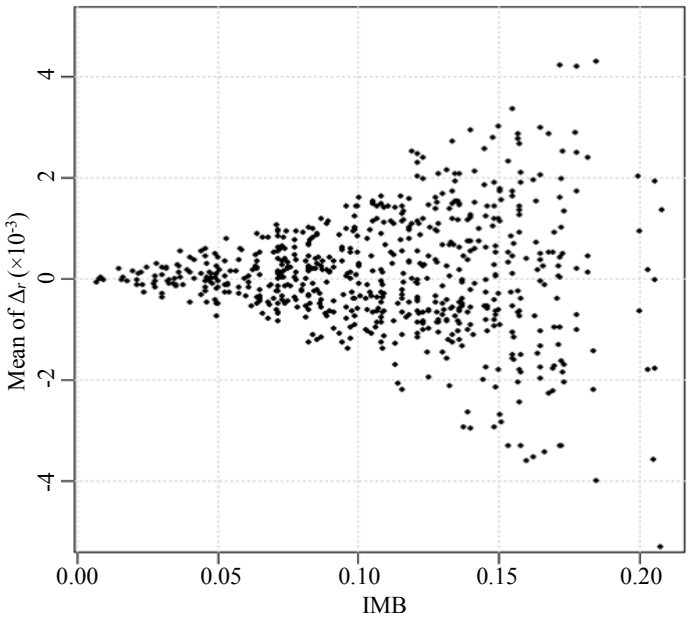

Figure 9.2 Conditional mean of Δ_r as a function of imbalance IMB; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20.

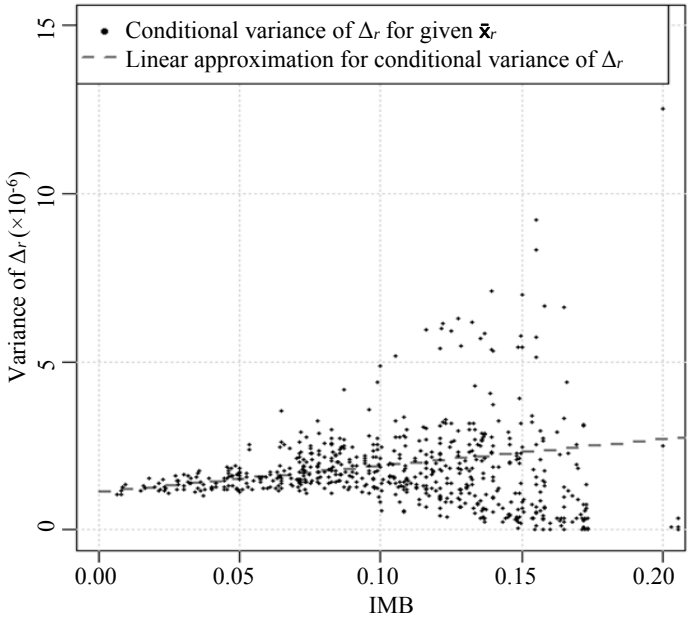

Figure 9.3 Conditional variance of Δ_r as a function of imbalance IMB; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20.

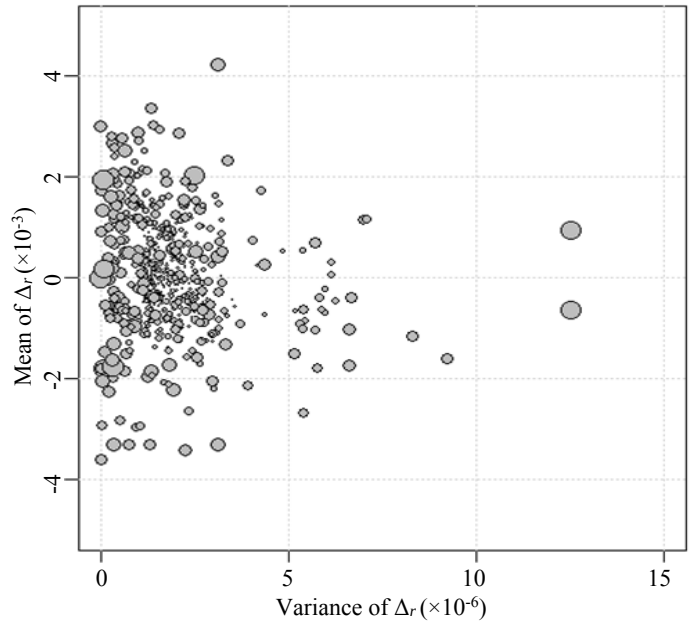

Figure 9.4 Plot of conditional mean of Δ_r against conditional variance of Δ_r; $\mathbf{x}_k$ is a non-group vector of dimension 5; response sets r of fixed size 12 from a fixed sample of size 20. Dot size proportional to imbalance squared.

10 Discussion

We comment on several issues arising and indicate limitations of our study.

1. Choice of variables for the auxiliary vector. The auxiliary variables for the vector $\mathbf{x}$ is treated as a fixed choice in this article. That choice is important when a perhaps large supply of such variables is available. Which ones should be chosen to meet the ultimate objective, which is best possible accuracy in the estimates? Result 1 shows that in the group vector case two factors are important for S_Δ^2 (which determines the conditional variance of CAL): The response imbalance IMB and the variance S_y^2 of the survey variable y. The fact that S_Δ^2 is (approximately) linearly decreasing with IMB gives incentive to try to reduce IMB in data collection. But allowing more variables in $\mathbf{x}$ increases IMB (because agreement is sought on more $\mathbf{x}$-means). As for the y-variance S_y^2, the trend is the opposite. By (7.1), S_y^2 is an averaged residual variance around group means; allowing additional variables in $\mathbf{x}$ will, especially if they explain y well, reduce S_y^2. The two factors work in opposite directions: More auxiliary variables give greater IMB but lower y-variance. It suggests a possible trade-off, a question not examined in this article. A particularity of a group vector $\mathbf{x}$ plays a role: When more categorical variables enter, the vector dimension grows in multiplicative bounds. The risk of small or empty cells restricts the expansion. To illustrate, if $\mathbf{x} = (sex \times education \times age)$ of dimension $J = 2 \times 3 \times 4 = 24$ is expanded to also include *occupation* with 4 categories, in completely crossed fashion, the new dimension (equal to the new number of groups) is $J = 24 \times 4 = 96$. In principle, S_y^2 decreases, but risk of small cells is a good reason to abstain from completely crossing all the variables and instead involve them in a non-group $\mathbf{x}$-vector. That case is addressed in Result 2, which says that if $\mathbf{x}$ explains y well, then σ_ε^2 is small and will give a desired low variance for Δ_r.

2. Auxiliary information at different levels. In this article, the imbalance IMB and the calibration estimator $\hat{Y}_{CAL}$ use the same $\mathbf{x}$-vector, and more particularly one that has auxiliary data for the sample units only. It is a realistic case. But in more general formulations, the data collection would use a monitoring vector $\mathbf{x}_{MV}$ possibly different from the calibration vector $\mathbf{x}_{CAL}$ used later in the estimation. The first is an instrument to get low imbalance IMB in the response, the second serves to get good calibrated weights for $\hat{Y}_{CAL}$. One reason why $\mathbf{x}_{MV}$ and $\mathbf{x}_{CAL}$ may differ in practice is that auxiliary variables for the estimation may be updated versions of the same variables available in the data collection. There may be other reasons to choose $\mathbf{x}_{MV}$ and $\mathbf{x}_{CAL}$ to be different. Also, they can contain information (if available) at the population level. Extensions of our approach to such situations are possible.

3. Uncertain benefit from reduced imbalance. Schouten et al. (2014) find evidence that balancing response reduces bias. We also find that there is incentive to strive, in data collection, for an ultimate response set with low imbalance IMB. As Results 1 and 2 show theoretically, and as test situations 1 and 2 confirm empirically, low imbalance gives a deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N}\Delta_r$ with zero or almost zero expected value and a small variance. This is our protection against large bias. If IMB were to increase, the variance

tends to increase. The zero expected value of the deviation $\hat{Y}_{CAL} - \hat{Y}_{FUL}$ is an average property. There is no guarantee that the deviation is small for any particular response r with low IMB.

4. Perfect balance does not eliminate the bias. Zero imbalance, $IMB = 0$, implies an equality of means for response and full sample, $\overline{\mathbf{x}}_r = \overline{\mathbf{x}}_s$. If that perfect balance were achieved, the bias adjustment term in (5.2) would be zero; the calibration (CAL) estimator and the expansion (EXP) estimator are identically equal. One can say that if perfect balance is achieved, the power of the auxiliary vector is exhausted, not in its potential for explaining the study variable, but in its potential for distancing itself from the crude EXP estimator, which, although it uses no auxiliary information at all, is as good as the otherwise better choice CAL. However, CAL ≡ EXP is still not ideal. As Result 1 shows, the variance of the CAL deviation is not near zero even if the imbalance IMB is near zero. Perfect balance does not eliminate the deviation of CAL, but small IMB protects against large deviation.

5. Practical implications. In this article we have primarily in mind surveys with a "substantial and non-eradicable nonresponse" that cannot realistically (under time and budget constraints for the survey) be brought to single-digit per cent levels even if large resources are spent. Surveys with 30 per cent or more nonresponse are common today. This is far from an ideal with near 100 per cent response, where imbalance and nonresponse would essentially cease to be issues; the EXP, CAL and FUL estimators would be close.

6. Directions for generalization. Results 1 and 2 show properties of the CAL deviation among response sets under a given formulation of the auxiliary vector. It would be desirable to generalize the results to other situations. Our proofs assume the existence of certain inverse matrices. Extensions to other cases would be possible with the aid of Moore-Penrose generalized inverse.

Acknowledgements

This work was supported by the Estonian Science Foundation grant 9127 and by the Institutional Research Funding IUT34-5 of Estonia. The authors gratefully acknowledge constructive comments from an Associate Editor and a Referee, both anonymous.

Appendix 1

Derivation of Result 1

We derive (7.2) to (7.4) under the conditions and notation in Section 7. The sample s is self-weighting, of size n, and $\mathbf{x}$ is a group vector of dimension J. We assume probability $\binom{n}{m}^{-1}$ for every response set r with fixed size m. We derive the expected value and the variance of $\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \overline{\mathbf{x}}_s = \sum_{j=1}^{J} W_{js} \overline{y}_{r_j} - \overline{y}_s$, where $W_{js} = n_j / n$, conditionally on fixed m and mean $\overline{\mathbf{x}}_r = (1/m)\left(m_1, \ldots, m_j, \ldots, m_J\right)$; $\sum_{j=1}^{J} m_j = m$.

Under that conditioning, $R = \prod_{j=1}^{J} \binom{n_j}{m_j}$ sets r have the same probability, where n_j is the size of sample group s_j; $\sum_{j=1}^{J} n_j = n$. This is identical to the probability structure for stratified simple random sampling of m_j from n_j in stratum j; $j = 1,\ldots,J$. Given m and $\bar{\mathbf{x}}_r$, the expected value and variance of $\bar{y}_{r_j}$ are, respectively, $\bar{y}_{s_j} = \sum_{s_j} y_k \big/ n_j$ and $\left(1/m_j - 1/n_j\right) S_{yj}^2$ with S_{yj}^2 given in (7.1). Thus $\bar{\Delta} = \sum_{j=1}^{J} W_{js} \bar{y}_{s_j} - \bar{y}_s = 0$, which proves (7.2), and $S_\Delta^2 = \sum_{j=1}^{J} W_{js}^2 \left(1/m_j - 1/n_j\right) S_{yj}^2$. Substituting $p_j = m_j / n_j$ and $p = m/n$, and using $S_y^2 = \sum_{j=1}^{J} W_{js} S_{yj}^2$ given in (7.1), we get

$$S_\Delta^2 = \frac{1}{n} \sum_{j=1}^{J} W_{js} \left(\frac{1}{p_j} - 1 \right) S_{yj}^2 = \left(\frac{1}{m} - \frac{1}{n} \right) S_y^2 + \frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) S_{yj}^2 . \tag{A.1}$$

This proves (7.3). To analyze the penalty term (second term on right hand side) in (A.1), suppose that the p_j vary little only around the overall rate p. Then $\delta_j = p_j / p - 1$, $j = 1,\ldots,J$, are small quantities, and $1/p_j = 1/p\left(1+\delta_j\right) = (1/p)\left(1 - \delta_j + \delta_j^2 - \delta_j^3 + \ldots\right)$. Keeping terms to second order, $p/p_j - 1 \approx -\delta_j + \delta_j^2$. The penalty term is then approximated as

$$\frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) S_{yj}^2 \approx -\frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right) S_{yj}^2 + \frac{1}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right)^2 S_{yj}^2 . \tag{A.2}$$

Let us further assume that the group variances S_{yj}^2, $j = 1,\ldots,J$, vary little only around their weighted mean S_y^2. Approximating $S_{yj}^2 \approx S_y^2$ in (A.2) we get

$$\frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p}{p_j} - 1 \right) \approx -\frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right) + \frac{S_y^2}{m} \sum_{j=1}^{J} W_{js} \left(\frac{p_j}{p} - 1 \right)^2 .$$

Here the first term on the right hand side is zero. The second term, equal to $(IMB/p^2)\left(S_y^2/m\right)$ with IMB given in (3.3), becomes a second approximation for the penalty term in (A.1). Therefore, $S_\Delta^2 \approx (1/m - 1/n) S_y^2 + (IMB/p^2)\left(S_y^2/m\right)$. This gives the desired result (7.4).

Appendix 2

Comparing two quadratic forms

We compare the two quadratic forms in $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$, Q_r and Q_s defined in (3.1), and justify the approximation $Q_r \approx Q_s$ needed in the proof in Appendix 3 of Result 2. The respective weighting matrices, $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$, are positive definite. Therefore Q_r (or Q_s) can be equal to zero only under the perfect balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. Since $Q_r = Q_s$ for perfect balance, the continuity argument implies that $Q_r \approx Q_s$ for nearly balanced response sets. How close are they more generally?

The CAL estimator (5.1) uses the weight factors $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$, defined for all $k \in s$. Their link to Q_r is shown in the second and third expressions in (A.3) below. Consider also the factors $f_k = \bar{\mathbf{x}}_r' \boldsymbol{\Sigma}_s^{-1} \mathbf{x}_k$ for

$k \in s$. They are instrumental for Q_s, and for $IMB = P^2 Q_s$, as the last two expressions in (A.3) show. The following moments of g_k and f_k are verified with the aid of the $\mathbf{x}$-vector condition (2.2):

$$\bar{g}_r = 1, \mathrm{var}_r(g) = Q_r, \ \bar{g}_s = 1 + Q_r; \quad \bar{f}_s = 1, \mathrm{var}_s(f) = Q_s, \ \bar{f}_r = 1 + Q_s. \tag{A.3}$$

For g_k, the means are defined as $\bar{g}_s = \sum_s d_k g_k \big/ \sum_s d_k$, $\bar{g}_r = \sum_r d_k g_k \big/ \sum_r d_k$, and the variances are $\mathrm{var}_s(g) = \sum_s d_k (g_k - \bar{g}_s)^2 \big/ \sum_s d_k$, $\mathrm{var}_r(g) = \sum_r d_k (g_k - \bar{g}_r)^2 \big/ \sum_r d_k$. For the corresponding moments of f_k, replace g_k by f_k. The variances $\mathrm{var}_s(g)$ and $\mathrm{var}_r(f)$ do not have an equally transparent form and will be approximated. Another important property following from (2.2) is $\sum_s d_k f_k g_k \big/ \sum_s d_k = \sum_r d_k f_k g_k \big/ \sum_r d_k = 1$. Those equations and appropriate expressions in (A.3) give

$$\mathrm{cov}_s(f,g) = \sum_s d_k (f_k - \bar{f}_s)(g_k - \bar{g}_s) \big/ \sum_s d_k = 1 - \bar{f}_s \bar{g}_s = -Q_r,$$

$$\mathrm{cov}_r(f,g) = \sum_r d_k (f_k - \bar{f}_r)(g_k - \bar{g}_r) \big/ \sum_r d_k = 1 - \bar{f}_r \bar{g}_r = -Q_s.$$

Now use $\mathrm{cov}_s^2(f,g) \le \mathrm{var}_s(f)\,\mathrm{var}_s(g)$ and the analogous inequality where r replaces s. Using also $\mathrm{var}_s(f) = Q_s$ and $\mathrm{var}_r(g) = Q_r$ from (A.3), we get bounds for the ratio Q_r / Q_s:

$$\frac{Q_s}{\mathrm{var}_r(f)} \le \frac{Q_r}{Q_s} \le \frac{\mathrm{var}_s(g)}{Q_r}. \tag{A.4}$$

For more transparent upper and lower bounds, approximate the two variances in (A.4) by assuming that the coefficient of variation (standard deviation divided by mean) is approximately the same for the response r as for the sample s, and this for both f and g. This assumes a certain stability of the coefficient of variation. Then $\mathrm{var}_s(g) \approx (\bar{g}_s)^2 \mathrm{var}_r(g) \big/ (\bar{g}_r)^2 = (1 + Q_r)^2 Q_r$, so the upper bound in (A.4) is approximately $(1 + Q_r)^2 > 1$. Similarly, $\mathrm{var}_r(f) \approx (\bar{f}_r)^2 \mathrm{var}_s(f) \big/ (\bar{f}_s)^2 = (1 + Q_s)^2 Q_s$, which gives $(1 + Q_s)^{-2} < 1$ as an approximate lower bound in (A.4). The interval approximation for the ratio Q_r / Q_s is therefore

$$Q_r / Q_s \in \left((1 + Q_s)^{-2}, (1 + Q_r)^2 \right).$$

This is to illustrate that the ratio is not far from 1, because for most data both Q_s and Q_r are small compared with 1, Q_r usually the somewhat bigger. Empirical work suggests however that the approximate upper bound $(1 + Q_r)^2$ can often be too low.

Appendix 3

Derivation of Result 2

We derive the expressions in (8.2) under the stated conditions. The sizes of r and s are m and n, respectively; the response rate is $p = m/n$. The deviation of CAL from the unbiased FUL is $\hat{Y}_{CAL} - \hat{Y}_{FUL} = \hat{N} \Delta_r$ where

$$\Delta_r = \left(\mathbf{b}_r - \mathbf{b}_s\right)' \bar{\mathbf{x}}_s = \Sigma_r d_k g_k y_k \,/ \Sigma_r d_k - \Sigma_s d_k y_k \,/ \Sigma_s d_k$$

with $\mathbf{b}_r$ and $\mathbf{b}_s$ given by (4.1), and $g_k = \bar{\mathbf{x}}_s' \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_k$. Note that $\mathbf{b}_s' \bar{\mathbf{x}}_s = \bar{y}_s$ by (2.2). Now $\Sigma_r d_k g_k \mathbf{x}_k' \,/\, \Sigma_r d_k = \Sigma_s d_k \mathbf{x}_k' \,/\, \Sigma_s d_k = \bar{\mathbf{x}}_s'$. Post-multiply that equation by $\boldsymbol{\beta}$ and use the result to get $\Delta_r = \Sigma_r d_k g_k \left(y_k - \mathbf{x}_k'\boldsymbol{\beta}\right) \big/ \Sigma_r d_k - \Sigma_s d_k \left(y_k - \mathbf{x}_k'\boldsymbol{\beta}\right) \big/ \Sigma_s d_k$, which expresses Δ_r in terms of the residuals $\varepsilon_k = y_k - \mathbf{x}_k'\boldsymbol{\beta}$ of the model (8.1):

$$\Delta_r = \frac{\sum_r d_k g_k \varepsilon_k}{\sum_r d_k} - \frac{\sum_s d_k \varepsilon_k}{\sum_s d_k}.$$

Then use the model properties of ε_k in (8.1). From $E_\xi\left(\varepsilon_k \mid \mathbf{x}_k\right) = 0$ for all k it follows that $E_\xi\left(\Delta_r \mid \mathbf{X}, r, s\right) = 0$. To evaluate the variance, use $E_\xi\left(\varepsilon_k^2 \mid \mathbf{x}_k\right) = \sigma_\varepsilon^2$, for all $k \in s$, and $E_\xi\left(\varepsilon_k \varepsilon_\ell \mid \mathbf{x}_k, \mathbf{x}_\ell\right) = 0$, all $k \neq \ell \in s$. This gives

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k^2}{\left(\sum_r d_k\right)^2} + \sigma_\varepsilon^2 \frac{\sum_s d_k^2}{\left(\sum_s d_k\right)^2} - 2\sigma_\varepsilon^2 \frac{\sum_r d_k^2 g_k}{\left(\sum_r d_k\right)\left(\sum_s d_k\right)}.$$

Here the d_k cancel out, because constant. The first and second expressions in (A.3) hold for any d_k, in particular constant d_k, so we get $\sum_r g_k \,/ m = 1$ for the mean and $\sum_r g_k^2 \,/ m = Q_r + 1$ for variance plus squared mean. Therefore,

$$E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) = \left(\frac{1}{m}\left(1 + Q_r\right) + \frac{1}{n} - 2\frac{1}{n}\right)\sigma_\varepsilon^2 = \left(\frac{1}{m} - \frac{1}{n} + \frac{Q_r}{m}\right)\sigma_\varepsilon^2.$$

As a final step, use the approximation $Q_r \approx Q_s$ justified in Appendix 2, and $IMB = p^2 Q_s$. Then, as claimed in Result 2, $E_\xi\left(\Delta_r^2 \mid \mathbf{X}, r, s\right) \approx \left(1 - p + IMB / p^2\right)\left(\sigma_\varepsilon^2 / m\right)$.

References

Beaumont, J.-F., Bocci, C. and Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-622.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of nonresponse in households surveys*. New York: John Wiley & Sons, Inc.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

Couper, M.P., and Wagner, J. (2011). Using paradata and responsive design to manage survey nonresponse. Proceedings, 58th World Statistics Congress, International Statistical Institute.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling. The cube method. *Biometrika*, 91, 893-912.

Groves, R. (2006). Research synthesis: Nonresponse rates and nonresponse error in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Little, R.J.A., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31, 2, 161-168. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2005002/article/9046-eng.pdf.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design. With applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rao, R.S., Glickman, M.E. and Glynn, R.J. (2008). Stopping rules for surveys with multiple waves of nonrespondent follow-up. *Statistics in Medicine*, 27, 2196-2213.

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

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: John Wiley & Sons, Inc.

Särndal, C.-E., and Lundquist, P. (2014a). Accuracy in estimation with nonresponse: A function of degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology,* 2, 361-387.

Särndal, C.-E., and Lundquist, P. (2014b). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société Française de Statistique*, 155(4), 28-50.

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology,* 39, 1, 29-58. Paper available at http://www.statcan.gc.ca/pub/12-001-x/12-001-x2013001-eng.pdf.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology,* 35, 1, 101-113. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2009001/article/10887-eng.pdf.

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, 727-748.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias. Ph. D. Thesis, University of Michigan, Ann Arbor.

Wagner, J., and Raghunathan, T.E. (2010). A new stopping rule for surveys. *Statistics in Medicine*, 29, 1014-1024.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 239-262
Statistics Canada, Catalogue No. 12-001-X

Statistical inference based on judgment post-stratified samples in finite population

Omer Ozturk[1]

Abstract

This paper draws statistical inference for finite population mean based on judgment post stratified (JPS) samples. The JPS sample first selects a simple random sample and then stratifies the selected units into H judgment classes based on their relative positions (ranks) in a small set of size H. This leads to a sample with random sample sizes in judgment classes. Ranking process can be performed either using auxiliary variables or visual inspection to identify the ranks of the measured observations. The paper develops unbiased estimator and constructs confidence interval for population mean. Since judgment ranks are random variables, by conditioning on the measured observations we construct Rao-Blackwellized estimators for the population mean. The paper shows that Rao-Blackwellized estimators perform better than usual JPS estimators. The proposed estimators are applied to 2012 United States Department of Agriculture Census Data.

Key Words: Post stratified sample; Finite sample correction; Ranked set sample; Stratified sample; Rao-Blackwellized estimator.

1 Introduction

In many survey sampling studies, in addition to variable of interest, sampling frame has additional available auxiliary variables to improve the information content of a sample. These auxiliary variables have been successfully used to construct better estimators, such as ratio and regression estimators. These estimators usually require strong modeling assumptions between the auxiliary variable(s) and variable of interest. MacEachern, Stasny and Wolfe (2004) introduced judgment post-stratified (JPS) sample, and constructed estimators that require weaker modeling assumptions than the ratio and regression estimators.

A JPS sample selects a simple random sample of size n from a population and measures all selected units, $X_i;\ i = 1,\ldots,n.$ For each one of the measured unit, researcher selects additional $H-1$ units to form a set of size H. This set contains the measured unit X_i and the additionally selected $H-1$ units. Units in these sets are ranked from smallest to largest without a measurement and the rank of X_i is determined. The pairs $(X_i, R_i); i = 1,\ldots,n,$ are called a JPS sample. Ranking process in these sets can be performed either using visual inspection of the units or some available auxiliary variable. If the visual inspection is used, rankers should be blinded to actual values of X_i to avoid any bias. If the auxiliary variable is used, a monotonic relationship between the variable X and auxiliary variable is required. These assumptions are much weaker than the linearity assumption in regression and ratio estimators.

Ranking information in a JPS sample is used to induce a structure among measured observations by creating H judgment classes of similar units. The judgment class h, $h = 1,\ldots,H,$ contains all measured observations with judgment rank h. Since rank R_i provides information about the relative position of X_i among H units in a set, observations in judgment class h are stochastically larger than the observations in

1. Omer Ozturk, The Ohio State University, Department of Statistics, 1958 Neil Avenue, Columbus, OH, 43210, U.S.A. E-mail: omer@stat.osu.edu.

judgment class h', for $h' < h$. This induced structure increases the information content of the sample. One may also view a JPS sample as a stratified sample with H strata. In this case, the improved efficiency can be established from standard theory of stratified sampling in survey sampling designs.

In a JPS sample, ranks are determined after a simple random sample is selected. Thus, the number of observations, M_h, in judgment class h is a random variable. The joint distribution of $\mathbf{M} = (M_1, \ldots, M_H)$ is multinomial with parameters n and success probability vector $(1/H, \ldots, 1/H)$. Since $\mathbf{M}$ is a random variable, it is highly possible that $M_h = 0$ for some h when the sample size n is small. Statistical inference then should account for the impact of empty strata on the procedures.

In an infinite population setting, JPS sample has generated extensive research interests. For a tiny slice of literature, readers are referred to Frey and Feeman (2012, 2013), Frey and Ozturk (2011), Stokes, Wang and Chen (2007), Wang, Lim and Stokes (2008), Wang, Stokes, Lim and Chen (2006), Wang, Wang and Lim (2012), Ozturk (2013, 2014a, 2014b, 2015) and the references there in.

One way to avoid having random sample size M_h is to rank the units in each set before selecting a simple random sample from the population. In this case sampling design is called ranked set sample (RSS). Ranked set sampling is introduced in McIntyre (1952, 2005) to estimate the population mean in agricultural research. To construct an RSS sample of size n, researcher first determines the design parameters, set size H and the judgment class sample size vector $\mathbf{m} = (m_1, \ldots, m_H)$, where m_h is the required number of observations to be selected in judgment class h. Researcher next selects nH units at random from the population and divide them into n sets, each of size H. Units in each one of these sets are ranked and the h^{th} judgment order statistics is measured in m_h sets so that $\sum_{h=1}^{H} m_h = n$. The measured observations $X_{[h]j}$; $j = 1, \ldots, m_h$; $h = 1, \ldots, H$ are called an unbalanced ranked set sample, where $X_{[h]j}$ is judgment order statistics from a set of size H. If the judgment class sample sizes are all equal $m_h \equiv n/H$; $h = 1, \ldots, H$, the sample is called a balanced ranked set sample. If there is no ranking error, judgment order statistics become usual order statistics from a sample of size H. In this case, usual order statistic notation is used to denote the h^{th} order statistic, $X_{(h)j}$.

In recent years, there have been increased research activities in JPS and RSS sampling in a finite population setting. Patil, Sinha and Taillie (1995) used ranked set sample to estimate population mean for a population of size N when the sample is constructed without replacement. Deshpande, Frey and Ozturk (2006) expanded the without replacement policy in Patil et al. (1995) into three different designs, design-0, design-1 and design-2, and constructed confidence intervals for population quantiles. The design-0 constructs the sample by replacing all units back into the population prior to selection of the next set. Design-1 constructs the sample by replacing only the unmeasured units back into the population before selecting the next set. Design-2 constructs the sample by replacing none of the units back into population regardless of whether they were measured or not. Al-Saleh and Samawi (2007), Ozdemir and Gokpinar (2007 and 2008), Jafari Jozani and Johnson (2011, 2012), Gokpinar and Ozdemir (2010), Ozturk and Jafari Jozani (2013), and Frey (2011) computed inclusion probabilities and constructed Horvitz-Thompson type estimators for population mean and total for some variant of design-0, design-1 and design-2 samples based on ranked set samples.

Ozturk (2014a) combined ranking information from different sources in a ranked set sample, estimated the inclusion probabilities of population units and constructed estimator for population mean. For settings where population values of auxiliary variables are available, Ozturk (2016) used population ranks (global ranking information) of selected sample units to induce stronger structure in data to improve the information content of the sample. He showed that samples constructed based on global ranking information provides higher efficiencies. A comprehensive up to date literature review both in JPS and RSS can be found in recent review paper in Wolfe (2012).

In this paper, we consider with- and without-replacement sampling designs for JPS sampling in finite population setting. Section 2 provides detailed descriptions for the construction of the designs. For each design, we obtain the probability mass functions, means, variances and covariances of order statistics. These results are used to construct unbiased estimator for the population mean and unbiased estimators for the variance of sample means. Section 3 constructs Rao-Blackwellized estimators by conditioning on the measured observations $X_i;\ i = 1,\ldots,n.$ Section 4 provides empirical evidence for the new estimators. Section 5 applies the proposed procedures to 2012 United States Agricultural Census (USDA) data. Section 6 provides some concluding remarks. The proofs of the theorems are provided in Appendix.

2 Sampling designs and estimator

We consider a finite population of size N, $\mathcal{P} = \{u_1,\ldots,u_N\}$, where u_j is the j^{th} unit in the population. Let X be the variable of interest. The values of X on population units will be denoted with $x_1,\ldots,x_N$. Without loss of generality, we assume that the population values of random variable X are ordered, $x_1 < \ldots < x_N$, so that the population rank of the unit u_i with respect to variable X is i, $R\left(x_{u_i}\right) = s_{u_i} = i$, where s_{u_i} is the rank of x_{u_i} among N population units. In addition to the variable of interest X, we assume that there is an additional variable Y that has monotonic relationship with random variable X.

We consider two sampling designs, design-0 and design-2. Both designs select a simple random sample $U_S = \{u_{s_1},\ldots,u_{s_n}\}$ from population $\mathcal{P}$. Without loss of generality, the sample U_S will be identified with rank vector $S = \{s_1,\ldots,s_n\}$. The design-0 selects the units with replacement, but design-2 selects the units without replacement. All selected sample units are measured for the variable X. Throughout the paper, we call $\mathbf{X} = (X_1,\ldots,X_n)$ as a sample of size n, where we use the notational convenience $X_i = X_{s_i}$. It is clear that $X_i;\ i = 1,\ldots,n$ are all independent in design-0, but they are negatively correlated in design-2. For each measured unit u_{s_i} in the sample, we randomly select an additional $H-1$ units without replacement from the remaining population units to form n sets each of size H,

$$S_{i,H} = \{u_{s_i}, u_{t_1},\ldots,u_{t_{H-1}}\};\ \ s_i \neq t_h, u_{t_h} \in \mathcal{P};\ \ h = 1,\ldots,H-1;\ \ i = 1,\ldots,n.$$

Units in each set $S_{i,H}$ are ranked based on auxiliary variable Y and the rank of the measured unit u_{s_i}, R_i, among H units is determined. Our judgment post-stratified sample then consists of pairs (X_i, R_i), $i = 1,\ldots,n.$ In design-0, all unmeasured units in set $S_{i,H}$ are replaced in the population before constructing the next set. Hence the same unmeasured unit(s) can appear in more than one sets. In design-2, none of the

unmeasured unit is returned to the population before constructing the next set. Hence, all sets $S_{i,H}$ are disjoint.

One can interpret rank vector $\mathbf{R} = \{R_1, \ldots, R_n\}$ as a covariate that replaces similar units, units having the same ranks, in the same judgment class. A JPS sample provides extra information on the measured unit u_{s_i} in addition to the measured value x_i through its relative position (rank R_i) in the set $S_{i,H}$. The quality of information depends on the strength of the monotonic relationship between the X and Y variables. It is clear that if the ranks, $R_i, i = 1, \ldots, n,$ are ignored, the sample is reduced to a simple random sample.

Ranking scheme is called consistent if the same ranking procedure is used in all sets. Under a consistent ranking scheme, following equalities hold.

Lemma 1. *Let (X_i, R_i) be a JPS sample constructed with a consistent ranking scheme and set size H from population $\mathcal{P}$.*

i. For design$-r$, $r = 0, 2,$ we have

$$\sum_{h=1}^{H} P\left(X_{[h]} = x\right) = P\left(X_j = x \,\middle|\, R_j = h\right) = \sum_{h=1}^{H} P\left(X_{(h)} = x\right).$$

ii. For design-2, we have

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{[h]} = x, X_{[h']} = y\right) = \sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_j = x, X_t = y \,\middle|\, R_j = h, R_t = h'\right)$$
$$= \sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{(h)} = x, X_{(h')} = y\right), x \neq y.$$

Part (i) of Lemma 1 is given in Presnell and Bohn (1999) in an infinite population setting. In this paper, we use a consistent judgment ranking scheme unless stated otherwise. Conditional mean and variance of X_i given $R_j = h$ and the conditional covariance of X_j, X_t given that $R_j = h, R_t = h'$ will be denoted by

$$\mu_{[h]} = E\left(X_{[h]}\right) = E\left(X_i \,|\, R_j = h\right),$$

$$\sigma^2_{[h]} = \mathrm{Var}\left(X_{[h]}\right) = \mathrm{var}\left(X_i \,|\, R_j = h\right)$$

and

$$\sigma_{[h,h']} = \mathrm{cov}\left(X_{[h]}, X_{[h']}\right) = \mathrm{cov}\left(X_j, X_t \,|\, R_j = h, R_t = h'\right).$$

Under perfect ranking, the square brackets in these expressions will be replaced with round parentheses.

There is a clear difference between the JPS samples in design-0 and design-2. In design-0, the pairs, $(X_i, R_i);\ i = 1, \ldots, n,$ are mutually independent. In design-2, any two measured observations X_i and X_j are negatively correlated even though their ranks R_i and R_j are independent. Ranks are independent

because they are determined independently in different sets. We first investigate the distributional properties of random variables in a JPS sample from population $\mathcal{P}$.

Lemma 2. *Let (X_i, R_i) be a JPS sample under perfect ranking with set size H from population $\mathcal{P}$.*

(i) Conditional probability mass function of X_j given $R_j = h$ is

$$\beta(i;h) = P\left(X_j = x_i \,\middle|\, R_j = h\right) = \frac{\binom{i-1}{h-1}\binom{N-i}{H-h}}{\binom{N}{H}}, x_i \in \mathcal{P}$$

for both design-0 and design-2 settings.

(ii) Conditional probability mass functions of X_r and X_t given that $R_r = h$ and $R_t = h'$, $\beta(i, j; h, h') = P\left(X_r = x_i, X_t = x_j \,\middle|\, R_r = h, R_t = h'\right)$ are

$$\beta_0(i, j; h, h') = \beta(i;h)\,\beta(j;h'), \left(x_i, x_j\right) \in \mathcal{P}$$

for design-0 and

$$\beta_1(i, j; h, h') = \sum_{\lambda=0}^{j-i-1} \frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}\binom{j-1-h-\lambda}{h'-1}\binom{N-j-H+\lambda+h}{H-h'}}{\binom{N}{H}\binom{N-H}{H}}, \; i < j, \left(x_i, x_j\right) \in \mathcal{P}$$

for design-2.

(iii) Conditional mean and variance of X_j given its rank $R_j = h$ are

$$\mu_{(h)} = E\left(X_j \,\middle|\, R_j = h\right) = \sum_{i=1}^{N} x_i \beta(i,h)$$

$$\sigma^2_{(h)} = \mathrm{Var}\left(X_j \,\middle|\, R_j = h\right) = \sum_{i=1}^{N} x_i^2 \beta(i,h) - \mu^2_{(h)}$$

for both design-0 and design 1.

(iv) Conditional covariance of X_r, X_t given their ranks are

$$\mathrm{cov}\left(X_r, X_t \,\middle|\, R_r = h, R_t = h'\right) = \sigma_{(h,h')} = 0$$

for design-0 and

$$\sigma_{(h,h')} = \sum_{i=1}^{N}\sum_{i \neq j}^{N} x_i x_j \beta_1(i, j, h, h') - \sum_{i=1}^{N} x_i \beta(i,h) \sum_{i=1}^{N} x_i \beta(i,h')$$

for design-2.

The proofs of part (i) and (ii) of the above Lemma are given in Patil et al. (1995). The proofs of the other parts are trivial and omitted here.

Ranking process in a JPS sample leads to a multinomial random vector $\mathbf{M} = (M_1, \ldots, M_H)$, where M_h is the number of observations in judgment class (post strata) h, $h = 1, \ldots, H$. The marginal distribution of M_h follows a binomial distribution with parameter n and $1/H$. For notational convenience we use $I_h = I(M_h > 0)$ to denote the event that the judgment class h is nonempty and $d_n = \sum_{h=1}^{H} I_h$ to define the number of non-empty judgment classes in a JPS sample. In the following Lemma, we provide some useful preliminary results on the judgment class sample size vector, proof of which can be found in Dastbaravarde, Arghami and Sarmad (2016) and Ozturk (2014b).

Lemma 3. *Let $(X_i, R_i); i = 1, \ldots, n$, be a JPS sample constructed under a consistent ranking scheme with set size H from $\mathcal{P}$. The following equalities hold for both design-0 and design-2:*

(i) $E\left(\frac{I_1}{d_n}\right) = 1/H$.

(ii) $E\left(\frac{I_1^2}{d_n^2}\right) = \frac{1}{H^2}\sum_{k=1}^{H}\left(\frac{k}{H}\right)^{n-1}$.

(iii) $Var\left(\frac{I_1}{d_n}\right) = \frac{1}{H^2}\sum_{k=1}^{H-1}\left(\frac{k}{H}\right)^{n-1}$.

(iv) $cov\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right) = -\frac{1}{H-1}Var\left(\frac{I_1}{d_n}\right)$.

(v) $E\left(\frac{I_1^2}{M_1 d_n^2}\right) = \frac{1}{H^n}\left\{\frac{1}{n} + \sum_{k=2}^{H}\sum_{j=1}^{k-1}\sum_{m_h=1}^{n-k+1}\frac{(-1)^{j-1}}{k^2 m_h}\binom{H-1}{k-1}\binom{k-1}{j-1}\binom{n}{m_h}(k-j)^{n-m_h}\right\}$.

We now consider the estimation of the population mean. We use

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

to denote the mean and variance of the population $\mathcal{P}$, respectively. Let

$$\hat{\mu}_r = \sum_{h=1}^{H}\frac{I_h}{M_h d_n}\sum_{i=1}^{n} X_i I(R_i = h),\ r = 0, 2$$

be the estimator for population mean μ based on design-r, $r = 0, 2$, respectively. In these estimators, I_h, M_h and d_n are random variables. They are used to make a correction on the estimator to yield an unbiased estimator for μ when some judgment classes are empty. If the ranks are ignored in a JPS sample, it becomes a simple random sample based on design-0 or design-2. In this case, population mean μ is estimated by

$$\bar{X}_r = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad r = 0, 2$$

from design-0 or design-2 data, respectively.

Theorem 1. *Let* $(X_i, R_i);\ i = 1, \ldots, n,$ *be a JPS sample of set size* H *constructed under a consistent ranking scheme based on either design-0 or design-2 from the finite population* $\mathcal{P}$. *(i) The estimators* $\hat{\mu}_r$ *are unbiased for* μ. *(ii) the variances of estimators* $\hat{\mu}_r$ *are*

$$\sigma^2_{\hat{\mu}_0} = \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 + E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h-1}^{H}\sigma^2_{[h]}$$

for $r = 0$ *and*

$$\sigma^2_{\hat{\mu}_2} = \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 - \frac{1}{H-1}\left(\frac{1}{H} - E\left(I_1/d_n\right)^2\right)\frac{H^2\sigma^2}{N-1}$$
$$+ E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h}^{H}\sigma^2_{[h]} + \left(\frac{H}{H-1}E\left(I_1/d_n\right)^2 - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H(H-1)}\right)\sum_{h=1}^{H}\sigma_{[h,h]}$$

for $r = 2$.

All expected values in $\sigma^2_{\hat{\mu}_0}$ and $\sigma^2_{\hat{\mu}_2}$ are computed over random sample size vector $\mathbf{M}$. These expected values can easily be computed from Lemma 2 using simple R-functions. Estimators for the population mean based on a balanced ranked set sample in design-0 and design-2 settings are given by

$$\mu^*_r = \frac{1}{mH}\sum_{i=1}^{m}\sum_{h=1}^{H} X_{[h]i}, \quad r = 0, 2,$$

where $m = n/H$ is the cycle size. Since the observations in design-0 are all independent, the variance of μ^*_0 is the same as the variance of RSS sample mean in an infinite population. The variance of μ^*_2 is given in equation 4.5 in Patil et al. (1995). In terms of our notation, the variance of μ^*_2 is written as

$$\sigma^2_{\mu^*_2} = \frac{(N-1-n)}{n(N-1)}\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 - \frac{1}{nH}\sum_{h=1}^{H}\sigma_{[h,h]}. \tag{2.1}$$

We put the variance of $\hat{\mu}_2$ in a slightly different format to compare it with $\sigma^2_{\mu^*_2}$

$$\sigma^2_{\hat{\mu}_2} = \sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2\left\{\frac{H}{H-1}\mathrm{Var}\left(I_1/d_n\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right)\right\}$$
$$+ \frac{H\sigma^2}{(N-1)(H-1)}\left\{(N-1)(H-1)E\left(\frac{I_1^2}{d_n^2 M_1}\right) - 1 + HE\left(\frac{I_1^2}{d_n^2}\right)\right\}$$
$$+ \left\{\frac{H}{H-1}E\left(I_1/d_n\right)^2 - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H(H-1)}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}. \tag{2.2}$$

One can easily see the impact of random sample size vector $\mathbf{M}$ on estimator $\hat{\mu}_2$ in a JPS sample by comparing equations (2.1) and (2.2). Expressions in curly brackets in equation (2.2) make corrections for the random sample sizes in JPS sample. For large population and sample sizes, $\sigma^2_{\mu^*_2}$, $\sigma^2_{\hat{\mu}_2}$ and $\sigma^2_{\hat{\mu}_0}$ reduce to simple forms.

Corollary 1. *Assume that* n *and* N *increases in such a way that the ratio of* n/N *approaches to a limit at* f, $\lim_{n\to\infty}(n/N) = f$.

(i) If $f > 0$, *the variances* $\sigma^2_{\hat{\mu}_2}$, $\sigma^2_{\mu_2^*}$ *and* $\sigma^2_{\hat{\mu}_0}$ *converge to two simple forms*

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\mu_2^*} = (1-f)\sigma^2 - \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 - \sum_{h=1}^{H}\sigma_{h,h}$$

and

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0} = \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]} = \sigma^2 - \frac{1}{H}\sum_{h=1}^{H}\left(\mu_h - \mu\right)^2.$$

(ii) If $f = 0$, $\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\mu_2^*} = \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0} = \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]}$ *which is the same as the variance of sample mean of a ranked set sample in infinite population setting.*

(iii) If f *is strictly positive, then* $\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_{RSS}} < \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0}$.

The part (iii) of the corollary indicates that when sample and population sizes grow at certain rate, the variances of sample means of JPS and RSS samples in finite population setting are always smaller than the variance of the same estimator in infinite population setting. This efficiency improvement is due to the negative correlation between X_i and X_j in without replacement sampling designs.

We now construct unbiased estimators for $\sigma^2_{\hat{\mu}_0}$, $\sigma^2_{\hat{\mu}_2}$ and $\sigma^2_{\mu_2^*}$. We first rewrite the estimators $\sigma^2_{\hat{\mu}_2}$ and $\sigma^2_{\mu_2^*}$ in slightly different forms

$$\begin{aligned}
\sigma^2_{\hat{\mu}_2} &= \left\{\frac{1}{H(H-1)} + E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right)\right\}\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} \\
&\quad + \frac{H^2\sigma^2}{H-1}\left\{\mathrm{Var}\left(\frac{I_1}{d_n}\right) - \frac{1}{N-1}\left\{\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right\}\right\} \\
&= C_1(n,H)\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} + C_2(n,H,N)\frac{H^2\sigma^2}{H-1} \qquad (2.3) \\
\sigma^2_{\mu_2^*} &= \frac{1}{Hn}\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} - \frac{\sigma^2}{(N-1)}.
\end{aligned}$$

In equation (2.3), it is clear that the coefficients $C_1(n,H)$ and $C_2(n,H,N)$ are known quantities for given values of sample size n and set size H. Let

$$\begin{aligned}
T_1 &= \frac{1}{E\left(\frac{I_1 I_2}{d_n^2}\right)}\sum_{h=1}^{H}\sum_{h\neq h'}^{H}\frac{I_h I_{h'}}{M_h M_{h'} d_n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_i - X_j\right)^2 I\left(R_i = h\right) I\left(R_j = h'\right), \\
T_2 &= \sum_{h=1}^{H}\frac{HI_h^*}{M_h d_n^*\left(M_h - 1\right)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(X_i - X_j\right)^2 I\left(R_i = h\right) I\left(R_j = h\right), \\
T_1^* &= \frac{1}{2m^2H^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(X_{[h]i} - X_{[h']j}\right)^2 \\
T_2^* &= \frac{1}{2m(m-1)H^2}\sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\left(X_{[h]i} - X_{[h]j}\right)^2
\end{aligned}$$

and

$$\hat{\sigma}^2_{SRS} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2,$$

where $I^*_h = I\left(M_h > 1\right)$, $d^*_n = \sum_{h=1}^{H} I^*_h$ and m is the cycle size in a balanced ranked set sample. From Lemma 3, one can easily establish that $E\left(I_1 I_2 / d^2_n\right) = \left(1/H - E\left(I_1/d_n\right)^2\right)/(H-1)$. Hence, T_1 is a statistic that depends only on the data. Note that $\hat{\sigma}^2_{SRS}$ is an unbiased estimator for σ^2. In the next theorem, we provide other unbiased estimators $(\bar{\sigma}^2)$ for σ^2 based on JPS and RSS samples in design-0 and design-2.

Theorem 2. *Let* $\left(X_i, R_i\right)$, $i = 1,\ldots,n$, *and* $X_{[h]i}$; $h = 1,\ldots,H$; $i = 1,\ldots,m$, *be JPS and RSS samples of set size* H, *respectively, both from the population* $\mathcal{P}$. *Unbiased estimators of* σ^2, $\sigma^2_{\hat{\mu}_0}$, $\sigma^2_{\hat{\mu}_2}$ *and* $\sigma^2_{\mu^*_2}$ *are given by, respectively,*

$$\bar{\sigma}^2 = \begin{cases} \left(T_1 + T_2\right)/\left(2H^2\right) & \textit{for design-0} \\ \dfrac{(N-1)\left(T_1+T_2\right)}{2NH^2} & \textit{for design-2} \\ \dfrac{\left(T^*_1 + T^*_2\right)(N-1)}{N} & \textit{for balanced RSS in design-2,} \\ T^*_1 + T^*_2 & \textit{for balanced RSS in design-0,} \end{cases} \tag{2.4}$$

$$\hat{\sigma}^2_{\hat{\mu}_0} = \frac{Var\left(I_1/d_n\right)}{2(H-1)}T_1 + \left\{E\left(\frac{I_1^2}{d_n^2 M_1}\right) - Var\left(\frac{I_1}{d_n}\right)\right\}\frac{T_2}{2}, \tag{2.5}$$

$$\hat{\sigma}^2_{\hat{\mu}_2} = C_1(n,H)T_2/2 + C_2(n,H,N)\frac{H^2\hat{\sigma}^2_{SRS}}{H-1}, \tag{2.6}$$

$$\bar{\sigma}^2_{\hat{\mu}_2} = C_1(n,H)T_2/2 + C_2(n,H,N)\frac{(N-1)\left(T_1+T_2\right)}{2N(H-1)}, \tag{2.7}$$

$$\hat{\sigma}^2_{\mu^*_2} = \frac{T^*_2}{m} - \frac{T^*_1 + T^*_2}{N}.$$

Note that both $\hat{\sigma}^2_{\hat{\mu}_2}$ and $\bar{\sigma}^2_{\hat{\mu}_2}$ are unbiased for $\sigma^2_{\hat{\mu}_2}$. The estimator $\bar{\sigma}^2$ is also unbiased for population variance σ^2 in RSS and JPS samples in design-0 and design-2. Theorem 2 indicates that all the variance estimators are unbiased for any sample size $n > 1$. We note that $E\left(I_1^2/d_n^2\right) \le 1/H^2$. By using this bound, one can show that $C_1(n,H) \ge 0$. On the other hand the coefficient $C_2(n,H,N)$ can be negative for some n, N and H. This rarely may lead to a negative value for $\hat{\sigma}^2_{\hat{\mu}_2}$. For negative $\hat{\sigma}^2_{\hat{\mu}_2}$, we propose a truncated estimator

$$\tilde{\sigma}^2_{\hat{\mu}_2} = \begin{cases} \hat{\sigma}^2_{\hat{\mu}_2} & \text{if } \hat{\sigma}^2_{\hat{\mu}_2} > 0 \\ C_1(n,H)T/2 & \text{if } \hat{\sigma}^2_{\hat{\mu}_2} \le 0. \end{cases} \tag{2.8}$$

This estimator is always positive and has very little bias. The values of $\overline{\sigma}^2_{\hat{\mu}_2}$ appears to be always positive based on our limited simulation study.

3 Rao-Blackwellized estimator

In this section, we construct estimators that improve the performance of the JPS estimators in the previous section. For a given simple random sample $X_{s_1},\ldots,X_{s_n}$, the sets $S_{j,H};\ j=1,\ldots,n,$ can be constructed over all possible matching of $n(H-1)$ additional units to n fully measured units. Each construction creates a new set of ranks hence a new estimate. We combine all these estimates by using Rao-Blackwell theorem. Let

$$\tilde{\mu}_r = E(\hat{\mu}\,|\,\mathbf{X}) = E\left\{\sum_{h=1}^{H}\frac{I_h}{M_h d_n}\sum_{j=1}^{n}X_j I\left(R_j=h\right)\Big|\,X_1,\ldots,X_n\right\}$$
$$= \sum_{i=1}^{n}X_i E\left[\sum_{h=1}^{H}\frac{I_h I\left(R_i=h\right)}{M_h d_n}\,\Big|\,\mathbf{X}\right] = \sum_{i=1}^{n}X_i a_{i|\mathbf{X}};\ r=0,2,$$

where

$$a_{i|\mathbf{X}} = E\left[\sum_{h=1}^{H}\frac{I_h I\left(R_i=h\right)}{M_h d_n}\,\Big|\,\mathbf{X}\right].$$

The expectation in $a_{i|\mathbf{X}}$ is taken over the conditional distributions of $R_j;\ j=1,\ldots,n,\ M_h;\ h=1,\ldots,N,$ and d_n given $\mathbf{X}$. We note that ranks, $R_j;\ j=1,\ldots,n,$ are assigned independently in each set $S_{j,H}$. Hence the joint distributions of $R_j;\ j=1,\ldots,n,$ given the measured observations $X_j;\ j=1,\ldots,n,$ are all independent

$$\alpha_{h_1,\ldots,h_n|\mathbf{X}} = P\left(R_1=h_1,\ldots,R_n=h_n\,|\,\mathbf{X}\right) = \prod_{j=1}^{n}P\left(R_j=h_j\,|\,\mathbf{X}\right),\ 1\le h_j\le H.$$

Assume that population ranks, $s_j;\ j=1,\ldots,n,$ of sample units are available, To construct the conditional distribution of $R_j=h_j$ given $X_j=x_{s_j}$, we first observe that

$$P\left(R_j=h_j, X_j=x_{s_j}\right) = \beta\left(s_j,h_j\right)/H \quad\text{and}\quad P\left(X_j=x_{s_j}\right)=1/N.$$

Using these joint and marginal probability mass functions, we write

$$\alpha_{h_j|s_j} = P\left(R_j=h_j\,|\,\mathbf{X}\right) = P\left(R_j=h_j\,|\,X_j=x_{s_j}\right) = \frac{\binom{s_j-1}{h_j-1}\binom{N-s_j}{H-h_j}}{\binom{N-1}{H-1}}, h_j=1,\ldots,H, x_{s_j}\in\mathcal{P}, \quad (3.1)$$

where x_{s_j} is the s_j^{th} smallest unit in the population. The evaluation of $a_{s_j|\mathbf{X}}$ in $\tilde{\mu}_r$ is computationally intensive. Even though the conditional distributions of rank R_j's are independent for given $\mathbf{X}$, they are not

identically distributed. Hence, the conditional distribution of judgment class sample size vector $\mathbf{M}$ given $\mathbf{X}$ does not have a multinomial distribution.

We now introduce an approximation to evaluate $\tilde{\mu}_r$. We first recognize that the conditional distribution of $R_j = h$ given $X_j = x_{s_j}$ is a hypergeometric distribution. Thus we can generate R_j from this hypergeometric distributions for given values of $\mathbf{X}$.

Algorithm 1.

I. Select an integer B. For $b = 1,\ldots,B$, generate $\mathbf{R}^{b,*} = \{R_1^{b,*},\ldots,R_n^{b,*}\}$ from $\alpha_{R_j \mid X_j = x_{s_j}}$ in equation (3.1).

II. Using $\mathbf{R}^{b,*}$ compute $I_h^{b,*} = I(M^{b,*} > 0)$, $M_h^{b,*} = \sum_{j=1}^{n} I\left(R_j^{b,*} = h\right)$, $d_n^{b,*} = \sum_{h=1}^{H} I_h^{b,*}$ and

$$a_i^{b,*} = \sum_{h=1}^{H} \frac{I^{b,*}}{M_h^{b,*} d_n^{b,*}} \sum_{h=1}^{H} I\left(R_i^{b,*} = h\right),\ i = 1,\ldots,n.$$

We approximate a_i with $\bar{a}_i^* = \sum_{b=1}^{B} a_i^{b,*}/B$, $i = 1,\ldots,n$. From law of large numbers $\bar{a}_i^*$ approaches to a_i as B gets large. Rao-Blackwellized estimator $\tilde{\mu}_r$ is then approximated by $\tilde{\mu}_r^* = \sum_{i=1}^{n} \bar{a}_i^* X_i$.

If the population ranks of sample units are not available, Algorithm 1 may not be usable. In this case, we use the collection of all unmeasured units to construct Rao-Blackwellized estimators. Let $\mathbf{Y}^{\mathrm{T}} = \left(Y_1,\ldots,Y_{n(H-1)}\right)$ be the auxiliary variables on unmeasured random variables. We use the following algorithm to approximate the Rao-Blackwellized estimators.

Algorithm 2. *For $b = 1,\ldots,B$, repeat the steps I-IV.*

I. Perform a random permutation on the entries of vector $\mathbf{Y}$ to obtain $\mathbf{Y}_b = permute(\mathbf{Y})$.

II. Divide the entries of $\mathbf{Y}_b$ into n sets, each of size $H-1$.

III. Match these n sets of size $H-1$ with n Y-values of the measured units to form n sets, each of size H. Obtain the rank of the X measurement from the rank of corresponding Y value in each set, $\mathbf{R}_b^* = \left(R_{b,1}^*,\ldots,R_{b,n}^*\right)$.

IV. Using $\mathbf{R}_b^*$ compute $I_h^{b,*} = I(M^{b,*} > 0)$, $M_h^{b,*} = \sum_{j=1}^{n} I\left(R_{b,j}^* = h\right)$, $d_n^{b,*} = \sum_{h=1}^{H} I_h^{b,*}$ and

$$a_i^{b,*} = \sum_{h=1}^{H} \frac{I^{b,*}}{M_h^{b,*} d_n^{b,*}} \sum_{h=1}^{H} I\left(R_i^{b,*} = h\right),\ i = 1,\ldots,n.$$

V. Compute the Rao-Blackwellized estimator $\tilde{\mu}_r = \sum_{i=1}^{n} \bar{a}_i^* X_i$, $r = 0, 2$, where $\bar{a}_i^* = \sum_{b=1}^{B} a_i^{b,*}/B$.

Even though large values of B provides better approximation to Rao-Blackwellized estimators, it may require additional computational effort and may not be feasible in practice. On the other hand, even small values of B, such as $B = 5$ could provide a significant improvement.

We now consider constructing estimators for the variance of Rao-Blackwellized estimators. Obtaining analytic expressions for the variances of $\tilde{\mu}_r$ is a challenge. Difficulty arises from the fact that there is no analytic expressions for the computation of $a_{i|\mathbf{X}}$; $i = 1,\ldots,n$. We then appeal to a bootstrap procedure to

compute the variance of $\tilde{\mu}_r$. Bootstrap estimators can be constructed from a plug-in method. Let θ be a statistical functional $\theta = T(\mathcal{P})$, where $\mathcal{P}$ is the finite population. The bootstrap estimate of θ can be obtained from $\hat{\theta} = T(\hat{\mathcal{P}})$, where $\hat{\mathcal{P}}$ is the empirical population. In finite population setting, the construction of the empirical population plays an important role to preserve the without replacement policies of bootstrap samples. Let K be the integer part of N/n and $k = N - Kn$. We construct $\hat{\mathcal{P}}$ with

$$\hat{\mathcal{P}} = \left\{ \underbrace{\mathbf{X}, \ldots, \mathbf{X}}_{K \text{ times}}, X_{t_1}, \ldots, X_{t_k} \right\},$$

where $X_{t_j}; \quad j = 1, \ldots, k,$ are selected at random from $\mathbf{X} = \{X_{s_1}, \ldots, X_{s_n}\}$. With this construction, the population size, N, is the same in both $\hat{\mathcal{P}}$ and $\mathcal{P}$. We generate bootstrap re-samples $\mathbf{X}^{*,1} = \{X^{*,1}_{s_1}, \ldots, X^{*,1}_{s_n}\}$ from $\hat{\mathcal{P}}$ with replacement for design-0 and without replacement for design-2. To construct the bootstrap distribution of the estimator $\tilde{\mu}_r$, we generate re-samples $\mathbf{X}^{*,c}, c = 1, \ldots, C$ and compute

$$\tilde{\mu}^{*,c}_r = \sum_{i=1}^{n} \overline{a}^*_i X^{*,c}_{s_i}, \quad c = 1, \ldots, C$$

from Algorithm 1 or 2. The bootstrap variance estimate of $\tilde{\mu}_r$ is then obtained from

$$\hat{\sigma}^2_{\tilde{\mu}_r} = \frac{1}{C-1} \sum_{c=1}^{C} \{\tilde{\mu}^{*,c}_r - \tilde{\mu}^*_r\}^2, \tag{3.2}$$

where $\tilde{\mu}^*_r$ is the mean of $\tilde{\mu}^{*,c}_r, c = 1, \ldots, C$.

A bootstrap $(1-\gamma)100\%$ percentile confidence interval for μ is constructed by $(L^{\gamma/2}_r, L^{1-\gamma/2}_r)$, where L^a_r is the a^{th} quantiles of $\tilde{\mu}^*_r$ satisfying $P(\tilde{\mu}^*_r \le L^a_r \,|\, \hat{\mathcal{P}})$ for $0 < a < 1$.

4 Empirical results

In this section, we look at the finite sample properties of the estimators in a small scale simulation study under wide ranges of simulation parameters. Data sets are generated from discrete normal and discrete shifted exponential populations for given population size N. The discrete populations are constructed from the quantile function

$$x_i = F^{-1}\left(\frac{i}{N+1}\right); \quad i = 1, \ldots, N, \tag{4.1}$$

where F is either normal or exponential cumulative distribution functions (CDF). For discrete normal population, we used location parameter 10 and scale parameter 4. For shifted discrete exponential population, we use the CDF of standard exponential distribution to generate x_i in equation (4.1) and then shift each x_i by adding 10. The population size is taken to be $N = 150$.

We used sample (n) and set size (H) to have integer values for n/H so that a balanced ranked set sample of size n can be created. Sample and set size combinations (n, H) are $(10, 2)$, $(15, 3)$, $(20, 4)$, $(25, 5)$. To control the quality of ranking information we used auxiliary variable Y, where $\rho = \text{cor}(X, Y)$ with $\rho = 1, 0.75$. The value of $\rho = 1$ yields perfect ranking and the value of $\rho = 0.75$ creates errors in

ranking. Simulation size is taken to be 3,000. Rao-Blackwellized estimators are computed from Algorithm 1 with $B = 50$ and bootstrap replication size 200.

The first part of the simulation investigates the efficiencies of the estimators and coverage probability of the confidence intervals of the population mean. All estimators are compared with design-2 Rao-Blackwellized estimators $(\tilde{\mu}_2)$. Let $D(\mathbf{X})$ be any one of the estimators introduced in Section 2 and 3. The relative efficiency of $D(.)$ with respect to $\tilde{\mu}_2$ is given by

$$R(D) = \frac{MSE(D)}{MSE(\tilde{\mu}_2)}, \tag{4.2}$$

where $MSE(D)$ is the estimated mean square error of estimator D. In equation (4.2), the value $R(D) > 1$ indicates that the estimator $\tilde{\mu}_2$ is more efficient than the estimator D.

We consider two types of confidence intervals for the population mean. Percentile confidence interval based on bootstrap distribution is given in Section 3. The coverage probabilities of these intervals will be labeled with $C^a(\tilde{\mu}_0)$ for design-0 and $C^a(\tilde{\mu}_2)$ for design-2. A second type of an approximate confidence interval can be constructed from standard theory. Note that we have unbiased estimators, $\hat{\sigma}_r$, for the variances of $\hat{\mu}_r$; $r = 0, 2$. A $100(1-\gamma)\%$ confidence interval for μ is then given by

$$\hat{\mu}_r \pm t_{n-1,1-\gamma/2}\hat{\sigma}_r; \ r = 0, 2,$$

where $t_{n-1,1-a}$ is the a^{th} upper quantile of the t-distribution with $n-1$ degrees of freedom. The coverage probabilities of these confidence intervals will be labeled as $C^b(\hat{\mu}_0)$ for design-0 and $C^b(\hat{\mu}_2)$ for design-2. Ahn, Lim and Wang (2014) suggested using $n-H$ degrees of freedom for the t-approximation. This selection may also work in JPS sampling in finite population setting with some increased variation due to unbalanced nature of a JPS sample. This line of work, on the other hand, is not persuaded in this paper because of the space limitation.

Table 4.1 presents the relative efficiencies of the estimators and the coverage probabilities of the confidence intervals for discrete normal populations. It is clear that Rao-Blackwellized design-2 estimator $(\tilde{\mu}_r)$ outperforms all the other estimators including RSS estimators. In general RSS estimators are more efficient than JPS estimators due to random judgment class sample size vector $\mathbf{M}$. This can be seen in Table 4.1 by looking at the ratio

$$\frac{R(\hat{\mu}_r)}{R(\mu_r^*)} = \frac{MSE(\hat{\mu}_r)}{MSE(\mu_r^*)}, \ r = 0, 2.$$

For $r = 0$, $\rho = 1$ and sample-set size combinations (n, H), $(10, 2), (15, 3), (20, 4), (25, 5)$, these ratios are $1.267(1.698/1.340)$, $1.491(2.117/1.419)$, $1.815(2.985/1.644)$, $2.391(3.479/1.455)$, respectively. It is obvious that ranked set sample estimator μ_0^* is more efficient that JPS estimator $\hat{\mu}_0$. This can be explained from the fact that RSS sample uses a constant (nonrandom) sample size vector $\mathbf{m} = (n_1, \ldots, n_H)$. Hence there is not extra variation due to randomness of $\mathbf{M}$ in JPS sample and this yields smaller variance for the estimator.

Table 4.1 (entries in columns $R(\mu_0^*)$ and $R(\mu_2^*)$) indicates that Rao-Blackwellized JPS estimators are better than RSS estimators. In this case, there is a clear difference between Rao-Blackwellized JPS

estimators and RSS sample estimators. In RSS sample, even though **m** is constant, ranking information (or rank R_i that belongs to each X_i) is obtained from a particular construction of n sets, each of size H. On the other hand, Rao-Blackwellized JPS estimators consider all possible constructions of n sets, each of size H. Hence, the content of ranking information is richer in a Rao-Blackwellized JPS sample than the content of ranking information of an RSS sample. This increased ranking information makes Rao-Blackwellized estimators superior to RSS estimators.

Table 4.1 also presents coverage probabilities of the confidence intervals. The coverage probabilities of bootstrap percentile confidence intervals are slightly lower than the nominal value 0.95. The coverage probabilities of the confidence intervals based on t-distribution are reasonably close to nominal coverage probability 0.95.

Table 4.1
Relative efficiencies of estimators and coverage probabilities of a 95% confidence interval of population mean. Data sets are generated from discrete normal population with mean μ = 10 and scale σ = 4

			Relative Efficiencies, $R(\bar{X}_0) = \mathrm{Var}(\bar{X}_0)/\mathrm{Var}(\tilde{\mu}_2)$							Coverage probabilities			
n	H	ρ	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu_0^*)$	$R(\mu_2^*)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	1.00	2.182	2.050	1.698	1.571	1.340	1.470	1.147	0.880	0.885	0.943	0.947
15	3	1.00	3.393	3.074	2.117	1.809	1.419	1.732	1.049	0.902	0.896	0.940	0.929
20	4	1.00	5.739	5.008	2.985	2.277	1.644	2.363	1.238	0.907	0.916	0.944	0.924
25	5	1.00	7.791	6.536	3.479	2.262	1.455	2.689	1.283	0.908	0.924	0.937	0.903
10	2	0.75	2.322	2.057	2.236	1.941	1.945	1.761	1.137	0.886	0.890	0.942	0.941
15	3	0.75	3.726	3.282	3.338	2.829	2.641	2.351	1.129	0.901	0.908	0.946	0.937
20	4	0.75	5.383	4.562	4.458	3.922	3.451	2.881	1.139	0.910	0.903	0.946	0.930
25	5	0.75	7.339	6.413	6.054	4.805	4.493	3.527	1.197	0.905	0.904	0.944	0.924

a: Coverage probabilities are computed from bootstrap percentile confidence interval.
b: Coverage probabilities are computed from $\hat{\mu}_r \pm t_{n-1,\,0.975}\hat{\sigma}_{\hat{\mu}_r}$, r = 0, 2.

Table 4.2 provides variance estimates of the mean estimators from simulation and the estimators in equations (2.5), (2.6), (2.8), and (3.2) in Sections 2 and 3. We already proved that the estimators $\hat{\sigma}^2_{\hat{\mu}_r}$, r = 0, 2, are unbiased. Entries for these variance estimators are very close to the corresponding values based on simulated variance estimates. The truncated variance estimator is almost identical to the un-truncated unbiased estimator. This shows that negative values happen rarely and there is not much difference between the truncated and un-truncated variance estimators. The bootstrap variance estimates of Rao-Blackwellized estimators are also very close to simulated variance estimates. Patterns similar to the ones we observed in Tables 4.1 and 4.2 also hold in Tables 4.3 and 4.4 for shifted exponential population.

Table 4.2
Variance estimate of the estimators. Data sets are generated from discrete normal population with mean μ = 10 and scale σ = 4

			Estimates from equations (2.5), (2.6), (2.8), (3.2)					Estimates from simulation			
n	H	ρ	$\hat{\sigma}^2_{\hat{\mu}_0}$	$\hat{\sigma}^2_{\hat{\mu}_2}$	$\tilde{\sigma}^2_{\tilde{\mu}_2}$	$\hat{\sigma}^2_{\tilde{\mu}_0}$	$\hat{\sigma}^2_{\tilde{\mu}_2}$	$V^a(\hat{\mu}_0)$	$V^a(\hat{\mu}_2)$	$V^a(\tilde{\mu}_0)$	$V^a(\tilde{\mu}_2)$
10	2	1.00	1.177	1.078	1.078	0.694	0.646	1.175	1.087	0.794	0.692
15	3	1.00	0.632	0.534	0.534	0.305	0.275	0.628	0.537	0.311	0.297
20	4	1.00	0.392	0.300	0.300	0.169	0.146	0.393	0.299	0.163	0.132
25	5	1.00	0.268	0.175	0.175	0.106	0.087	0.270	0.175	0.099	0.078
10	2	0.75	1.431	1.335	1.335	0.692	0.645	1.463	1.270	0.744	0.654
15	3	0.75	0.896	0.802	0.802	0.306	0.276	0.901	0.763	0.305	0.270
20	4	0.75	0.631	0.531	0.531	0.169	0.145	0.627	0.552	0.160	0.141
25	5	0.75	0.485	0.386	0.386	0.106	0.089	0.506	0.401	0.100	0.083

a: These variance estimates are obtained from simulation.

Table 4.3
Relative efficiencies of estimators and coverage probabilities of a 95% confidence interval of population mean. Data sets are generated from discrete shifted exponential population with scale $\sigma = 1$ and shift parameter 10

			Relative Efficiencies, $R(\bar{X}_0) = \mathrm{Var}(\bar{X}_0)/\mathrm{Var}(\tilde{\mu}_2)$							Coverage probabilities			
n	H	ρ	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu_0^*)$	$R(\mu_2^*)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	1.00	1.770	1.663	1.495	1.394	1.193	1.297	1.103	0.838	0.833	0.894	0.950
15	3	1.00	2.472	2.239	1.757	1.538	1.222	1.446	1.027	0.855	0.842	0.905	0.931
20	4	1.00	3.839	3.349	2.353	1.889	1.406	1.879	1.212	0.871	0.884	0.915	0.931
25	5	1.00	4.639	3.892	2.503	1.792	1.235	1.958	1.182	0.865	0.881	0.916	0.915
10	2	0.75	1.900	1.690	1.941	1.690	1.667	1.520	1.128	0.839	0.857	0.898	0.949
15	3	0.75	2.708	2.440	2.626	2.233	2.132	1.815	1.117	0.859	0.870	0.914	0.947
20	4	0.75	3.484	2.996	3.059	2.704	2.430	2.103	1.104	0.869	0.871	0.922	0.938
25	5	0.75	4.758	4.127	4.156	3.298	3.106	2.402	1.245	0.866	0.877	0.913	0.932

a : Coverage probabilities are computed from bootstrap percentile confidence interval.
b : Coverage probabilities are computed from $\hat{\mu}_r \pm t_{n-1,\,0.975}\hat{\sigma}_{\hat{\mu}_r}$, $r = 0, 1$.

Table 4.4
Variance estimate of the estimators. Data sets are generated from discrete shifted exponential population with scale $\sigma = 1$ and shift parameter 10

			Estimates from equations (2.5), (2.6), (2.8), (3.2)					Estimates from simulation			
n	H	ρ	$\hat{\sigma}^2_{\hat{\mu}_0}$	$\hat{\sigma}^2_{\hat{\mu}_2}$	$\tilde{\sigma}^2_{\tilde{\mu}_2}$	$\hat{\sigma}^2_{\tilde{\mu}_0}$	$\hat{\sigma}^2_{\tilde{\mu}_2}$	$V^a(\hat{\mu}_0)$	$V^a(\hat{\mu}_2)$	$V^a(\tilde{\mu}_0)$	$V^a(\tilde{\mu}_2)$
10	2	1.00	0.077	0.069	0.069	0.046	0.042	0.075	0.070	0.055	0.050
15	3	1.00	0.042	0.036	0.036	0.022	0.020	0.042	0.037	0.025	0.024
20	4	1.00	0.027	0.022	0.022	0.013	0.012	0.027	0.022	0.014	0.012
25	5	1.00	0.019	0.014	0.014	0.009	0.007	0.019	0.014	0.009	0.008
10	2	0.75	0.089	0.083	0.083	0.046	0.043	0.090	0.078	0.052	0.046
15	3	0.75	0.055	0.051	0.051	0.022	0.020	0.057	0.048	0.024	0.022
20	4	0.75	0.039	0.033	0.033	0.013	0.011	0.039	0.034	0.014	0.013
25	5	0.75	0.031	0.025	0.025	0.009	0.008	0.032	0.025	0.009	0.008

a : These variance estimates are obtained from simulation.

5 Example

In this section we apply the proposed estimators to estimate corn production in Ohio based on 2012 United States Department of Agriculture (USDA) census. The population consists of $N = 87$ counties in Ohio (One of the county is excluded from the population since census data did not have any entry for it). Variable of interest is the total corn production (X) in bushels. We use 2007 USDA census corn production (Y) as an auxiliary variable. Mean and standard deviation of corn production in 2012 are $\mu_X = 5{,}021{,}061$ and $\sigma_X = 3{,}983{,}560$ bushels, respectively. The correlation coefficient between X and Y is 0.963. Using this population, we performed another simulation study to estimate the corn production and constructed confidence intervals for the population mean. Samples are generated for sample and set size combinations $(n, H) = (10, 2), (15, 3), (20, 4)$. Simulation and bootstrap replications sizes are taken to be 3,000 and 200, respectively. Rao-Blackwellized estimators are computed based on 50 replications.

Relative efficiencies of the estimators with respect to $\tilde{\mu}_2$ and coverage probabilities of the confidence intervals are given in Table 5.1. Table 5.1 indicates that Rao-Blackwellized design-2 estimators outperforms all the other estimators we considered. Coverage probabilities appear to be slightly smaller than the nominal level 0.95.

Table 5.1
Relative efficiencies of estimators and coverage probabilities of a 95% confidence interval of population mean. The population is 87 Ohio counties. Variable of interest is corn production (X) in 2012. Auxiliary variable is corn production (Y) in 2007, μ_X = 5,021,061, σ_X = 3,983,560, $\mathrm{cor}(X,Y)$ = 0.963 and N = 87

		Relative Efficiencies, $R(\bar{X}_0) = \mathrm{Var}(\bar{X}_0)/\mathrm{Var}(\tilde{\mu}_2)$							Coverage probabilities			
n	H	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu_0^*)$	$R(\mu_2^*)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	2.301	1.981	1.829	1.448	1.468	1.280	1.181	0.883	0.896	0.924	0.925
15	3	3.745	3.188	2.353	1.612	1.994	1.454	1.200	0.907	0.919	0.940	0.907
20	4	5.707	4.402	2.901	1.624	2.476	1.143	1.341	0.920	0.920	0.946	0.873

a : Coverage probabilities are computed from bootstrap percentile confidence interval.
b : Coverage probabilities are computed from $\hat{\mu}_r \pm t_{n-1,\,0.975}\hat{\sigma}_{\hat{\mu}_r}$, $r = 0, 2$.

Table 5.2 presents the estimates of the standard deviation of the estimators of population mean from simulations and from analytic expression in equation (2.5), (2.6), (2.8), (3.2). It is again clear that estimates of the standard errors are reasonably close to the estimates from simulations. The standard deviation estimates of the estimators of the population total are obtained by multiplying the entries in Table 5.2 with the population size $N = 87$.

Table 5.2
Estimates of the standard deviation of the estimators from 2012 USDA census. The population is 87 Ohio counties. Variable of interest is corn production (X) in 2012. Auxiliary variable is corn production (Y) in 2007, μ_X = 5,021,061, σ_X = 3,983,560, $\mathrm{cor}(X,Y)$ = 0.963 and N = 87

		Estimates from equations (2.5), (2.6), (2.8), (3.2)					Estimates from simulation			
n	H	$\hat{\sigma}_{\hat{\mu}_0}$	$\hat{\sigma}_{\hat{\mu}_2}$	$\tilde{\sigma}_{\hat{\mu}_2}$	$\hat{\sigma}_{\tilde{\mu}_0}$	$\hat{\sigma}_{\tilde{\mu}_2}$	$\sqrt{V^a(\hat{\mu}_0)}$	$\sqrt{V^a(\hat{\mu}_2)}$	$\sqrt{V^a(\tilde{\mu}_0)}$	$\sqrt{V^a(\tilde{\mu}_2)}$
10	2	1,108,818.7	1,027,289.0	1,027,717.4	883,847.8	833,711.4	1,156,300.5	1,028,629.9	929,090.9	854,940.3
15	3	815,371.3	687,605.0	689,118.9	602,682.4	545,000.2	810,156.1	670,521.9	578,608.4	528,146.5
20	4	652,734.4	472,231.5	477,888.6	454,368.3	392,990.3	638,755.1	478,007.6	434,365.0	375,040.7

a : These variance estimates are obtained from simulation.

6 Concluding remark

We have developed two sampling designs for judgment post stratified samples in a finite population setting. The designs are constructed with two levels of without-replacement policies, design-0 and design-2. Design-0 is constructed by replacing all measured and unmeasured units in a set back in population before selecting the next units. Design-2 is constructed by using without-replacement policy on all units regardless of the measurement status. Hence, random variables in design-0 are independent, but random variables in design-2 are negatively correlated. In these designs, measured observations are ranked and stratified into H judgment classes after a simple random sample is collected. Using this ranking information, we construct unbiased estimators for the population mean and the variance of these unbiased estimators of population mean. We showed that the new estimators based on level-2 design outperform simple random sample mean, design-0 JPS sample mean and ranked set sample mean estimators. Main focus of this paper was on the estimation of populations mean, but the results also apply to estimation of population total with a minor adjustment in notation.

Post stratification creates random ranks and random number of observations in judgment classes. By conditioning on the measured values in the sample, we construct Rao-Blackwellized estimators by

computing the conditional expected value of the estimators over all possible values of random ranks. Rao-Blackwellized estimators are unbiased and more efficient than unconditional estimators. We construct finite sample bootstrap inference for the population mean based on all proposed estimators. The new sampling designs and estimator are applied to 2012 USDA census data to show that they are viable sampling designs and estimators in survey sampling studies.

In one of our current projects, we extend these design-0 and design-2 to two-stage sampling where primary and secondary sampling units constitute two finite populations. In this case, we expect that some interesting optimization problem will arise related the selection of sample sizes and design-0 and design-2 in stage I and II sampling.

Appendix

Proof of Lemma 1: The proof of part (i) is given in Presnell and Bohn (1999) in an infinite population setting. The proof is essentially the same in finite population setting.

For the proof of part (ii), we consider the joint probability mass function of X_j and X_t given their judgment ranks $R_j = h$ and $R_t = h'$

$$P\left(X_j = x, X_t = y \,|\, R_j = h, R_t = h'\right) = P\left(X_{[h]} = x, X_{[h']} = y\right);\ x \neq y, (x,y) \in \mathcal{P}.$$

For $h = 1,\ldots,H$, let $J_h = i$ be the event that i^{th} order statistic in a set of size H is judged to be the h^{th} judgment order statistics. Using this notation, we write

$$P\left(X_{[h]} = x, X_{[h']} = y\right) = \sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y, J_h = i, J_{h'} = k\right);\ x \neq y, (x,y) \in \mathcal{P}. \qquad \text{(A.1)}$$

For each i, the events $\{J_h = i\};\ h = 1,\ldots,H$, partition the sample space of random variable R_j because the i^{th} order statistic is assigned to one and only one judgment rank. Since rank R_t is obtained independently from another disjoint set using the same ranking procedure, the events $\{J_{h'} = k\};\ h' = 1,\ldots,H$, also partition the sample space. Hence, we write

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y, J_h = i, J_{h'} = k\right) = P\left(X_{(i)} = x, X_{(k)} = y\right);\ x \neq y, (x,y) \in \mathcal{P}. \qquad \text{(A.2)}$$

Combining equations (A.1) and (A.2), we obtain

$$\begin{aligned}\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{[h]} = x, X_{[h']} = y\right) &= \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y;\ J_h = i, J_{h'} = k\right) \\ &= \sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y\right);\ x \neq y, (x,y) \in \mathcal{P}.\end{aligned}$$

This completes the proof.

Proof of Theorem 1:

(i) We use the conditional expectation to write

$$E(\hat{\mu}_r) = \sum_{h=1}^{H} E\left(\frac{I_h}{d_n M_h}\right) \sum_{i=1}^{n} E\left(X_i I(R_i = h) \middle| R_i = h\right) = \sum_{h=1}^{H} E\left(\frac{I_h M_h}{d_n M_h}\right) EX_{[h]}.$$

Random variables I_h / d_n; $h = 1, \ldots, H$, are identically distributed. By using part (i) in Lemma 3, the above expectation can be written

$$E(\hat{\mu}_r) = E\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{n} EX_{[h]} = \frac{1}{H} \sum_{h=1}^{H} \mu_{[h]} = \frac{1}{H} \sum_{h=1}^{H} \mu_{(h)} = \mu, r = 0, 2$$

which completes the proof. The last two equalities in the above equation follows from part (i) of Lemma 1.

(ii) For the the proof of $\mathrm{Var}(\hat{\mu}_0)$, we use conditional variance

$$\mathrm{Var}(\hat{\mu}_0) = E\{\mathrm{Var}(\hat{\mu}_0 | \mathbf{R})\} + \mathrm{Var}\{E(\hat{\mu}_0 | \mathbf{R})\}.$$

Since X_j are selected with replacement, they are all independent. Hence, the first term in the above equation yields

$$E\{\mathrm{Var}(\hat{\mu}_0 | \mathbf{R})\} = \sum_{h=1}^{H} E\left(\frac{I_h^2 M_h}{d_n^2 M_h^2}\right) \sigma_{[h]}^2 = E\left(\frac{I_1^2}{d_n^2 M_1}\right) \sum_{h=1}^{H} \sigma_{[h]}^2.$$

We now consider

$$\begin{aligned}
\mathrm{Var}\{E(\hat{\mu}_0 | \mathbf{R})\} &= \sum_{h=1}^{H} \mathrm{Var}\left(\frac{I_h}{d_n}\right) \mu_{[h]}^2 + \sum_{h=1}^{H} \sum_{h' \neq h}^{H} \mathrm{Cov}\left(\frac{I_h}{d_n}, \frac{I_{h'}}{d_n}\right) \mu_{[h]} \mu_{[h']} \\
&= \mathrm{Var}\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{H} \mu_{[h]}^2 + \mathrm{Cov}\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right) \sum_{h=1}^{H} \sum_{h' \neq h}^{H} \mu_{[h]} \mu_{[h']} \\
&= \frac{H}{H-1} \mathrm{Var}\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{H} \left(\mu_{[h]} - \mu\right)^2
\end{aligned}$$

which completes the proof of $\mathrm{Var}(\hat{\mu}_0)$. The last equality is obtained by using part (iv) of Lemma 3.

For the proof of $\mathrm{Var}(\hat{\mu}_2)$, we again consider the conditional variance of $\hat{\mu}_2$ given the ranks

$$\mathrm{Var}(\hat{\mu}_2) = E(\mathrm{Var}(\hat{\mu}_2 | \mathbf{R})) + \mathrm{Var}(E(\hat{\mu}_2 | \mathbf{R})).$$

The second term in the right hand side of the above equation can be written

$$\begin{aligned}
\mathrm{Var}(E(\hat{\mu}_2 | \mathbf{R})) &= \mathrm{Var}\left(\sum_{h=1}^{h} \frac{\mu_{[h]} I_h}{d_n}\right) \\
&= \sum_{h=1}^{H} \mathrm{Var}\left(\frac{I_h}{d_n}\right) \mu_{[h]}^2 + \sum_{h=1}^{H} \sum_{h' \neq h}^{H} \mathrm{Cov}\left(\frac{I_h}{d_n}, \frac{I_{h'}}{d_n}\right) \mu_{[h]} \mu_{[h']} \\
&= \mathrm{Var}\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{H} \mu_{[h]}^2 + \mathrm{Cov}\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right) \sum_{h=1}^{H} \sum_{h' \neq h}^{H} \mu_{[h]} \mu_{[h']}.
\end{aligned}$$

Note that $\mathrm{Cov}(I_1/d_n, I_2/d_n) = -\mathrm{Var}(I_1/d_n)/(H-1)$ in part (iv) of Lemma 3. Using this equality in the above equation, we obtain

$$\mathrm{Var}(E(\hat{\mu}_2 | \mathbf{R})) = \frac{H}{H-1} \mathrm{Var}\left(\frac{I_1}{d_n}\right) \sum_{h=1}^{H} \left(\mu_{[h]} - \mu\right)^2. \qquad \text{(A.3)}$$

We now consider $\mathrm{Var}\left(\hat{\mu}_2 \,|\mathbf{R}\right)$

$$\begin{aligned}\mathrm{Var}\left(\hat{\mu}_2 \,|\mathbf{R}\right) &= \mathrm{Var}\left\{\sum_{h=1}^{H}\frac{I_h}{d_n M_h}\sum_{j=1}^{n} X_j I\left(R_j = h\right)\Big|\mathbf{R}\right\} \\ &= \sum_{h=1}^{H}\frac{I_h^2}{d_n^2 M_h^2}\left(M_h\sigma_{[h]}^2 + M_h\left(M_h-1\right)\sigma_{[h,h]}\right) + \sum_{h}^{H}\sum_{h'\neq h}^{H}\frac{I_h I_{h'}}{d_n^2 M_h M_{h'}} M_h M_{h'}\sigma_{[h,h']} \\ &= \sum_{h=1}^{H}\frac{I_h^2}{d_n^2 M_h}\sigma_{[h]}^2 + \sum_{h=1}^{H}\frac{I_h^2\left(M_h-1\right)}{d_n^2 M_h}\sigma_{[h,h]} + \sum_{h}^{H}\sum_{h'\neq h}^{H}\frac{I_h I_{h'}}{d_n^2}\sigma_{[h,h']}.\end{aligned}$$

It is clear that $I_h/\left(d_n M_h\right)$ and $I_h I_{h'}/d_n^2$, $h = 1,\ldots,H$, are identically distributed. Using the equalities below

$$\begin{aligned}E\left(\frac{\left(M_1-1\right)I_1^2}{d_n^2 M_1}\right) &= E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) \\ E\left(\frac{I_1 I_2}{d_n^2}\right) &= \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\end{aligned}$$

we write

$$\begin{aligned}E\left\{\mathrm{Var}\left(\hat{\mu}_2 \,|\mathbf{R}\right)\right\} &= E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 + E\left(\frac{I_1^2}{d_n^2}\right)\sum_{h=1}^{H}\sigma_{[h,h]} - E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h,h]} \\ &\quad + \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} \\ &= E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 + \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\sum_{h=1}^{H}\sum_{h=1}^{H}\sigma_{[h,h']} \\ &\quad + \left\{E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\right\}\sum_{h=1}^{H}\sigma_{[h,h]}. \qquad \text{(A.4)}\end{aligned}$$

We now show that $\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma_{[h,h']} = H^2\sigma^2/(N-1)$. Using part (ii) of Lemma 1, we first observe that

$$\begin{aligned}\sum_{h=1}^{H}\sum_{h'=1}^{H}E\left(X_{[h]}X_{[h']}\right) &= \sum_{h=1}^{H}\sum_{h'=1}^{H}E\left(X_{(h)}X_{(h')}\right) \\ &= 2\sum_{i=1}^{}\sum_{i<j}^{}x_i x_j\sum_{h=1}^{}\sum_{h'=1}^{}\beta\left(i,j,h,h'\right) \\ &= 2\sum_{i=1}^{}\sum_{j>i}^{}x_i x_j\left\{\sum_{h=1}^{H}\sum_{\lambda=0}^{j-i-1}\frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}}{\binom{N}{H}}\right. \\ &\qquad \left.\times\sum_{h'=1}^{H}\frac{\binom{j-1-h-\lambda}{h'-1}\binom{N-j-H+\lambda+h}{H-h'}}{\binom{N-H}{H}}\right\}.\end{aligned}$$

In the last sum of the above equation, we let $y = h' - 1$. After some simplification, we write

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) = 2\sum\sum_{i<j} x_i x_j \sum_{h=1}^{H}\sum_{\lambda=0}^{j-i-1} \frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}}{\binom{N}{H}} \frac{H}{N-H}$$

$$= \frac{H}{N-H}\sum\sum_{i\neq j} x_i x_j \sum_{h=1}^{H} \frac{\binom{i-1}{h-1}\binom{N-i-1}{H-h}}{\binom{N}{H}}.$$

Let $z = h - 1$. The above expression reduces to

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) = \frac{H}{N-H}\sum\sum_{i\neq j} x_i x_j \sum_{z=0}^{H-1} \frac{\binom{i-1}{z}\binom{N-i-1}{H-1-z}}{\frac{N(N-1)}{H(H-N)}\binom{N-1}{H-1}}$$

$$= \frac{H^2}{N(N-1)}\sum\sum_{i\neq j} x_i x_j = \frac{H^2}{N(N-1)}\left\{\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i}^{N} x_i^2\right\}.$$

Using the above equation, we conclude that

$$\sum\sum_{h=1}^{H}\sum_{h'=1} \sigma_{[h,h']} = \sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) - \sum\sum_{h=1}^{H}\sum_{h'=1} E\left(X_{[h]}EX_{[h']}\right)$$

$$= \frac{H^2}{N(N-1)}\left\{\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i}^{N} x_i^2\right\} - \frac{H^2}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2$$

$$= -\frac{H^2}{N-1}\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = \frac{-H^2}{N-1}\sigma^2. \qquad \text{(A.5)}$$

By inserting the above expression in equation (A.4), we write

$$E\left\{\mathrm{Var}\left(\hat{\mu}_2\right)\middle|\mathbf{R}\right\} = E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 - \frac{H^2\sigma^2}{(H-1)(N-1)}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)$$

$$+\left\{\frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H(H-1)}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}. \qquad \text{(A.6)}$$

We complete the proof by combining equations (A.3) and (A.6)

$$\mathrm{Var}\left(\hat{\mu}_2\right) = \frac{H}{H-1}\mathrm{Var}\left(I_1/d_n\right)\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 + E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2$$

$$-\frac{H^2\sigma^2}{(H-1)(N-1)}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)$$

$$+\left\{\frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{(H-1)H}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}.$$

Proof of Corollary 1: Proofs of (i) and (ii) are trivial. For the proof of (iii), we rewrite $\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2}$ as

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \frac{1}{H}\left(\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{[h,h]}\right) - f\sigma^2.$$

Using the equality (A.5), we write

$$\begin{aligned}
-\frac{1}{H}\sum_{h=1}^{H}\sigma_{h,h} - f\sigma^2 &\approx -\frac{NH}{NH^2}\sum_{h=1}^{H}\sigma_{[h,h]} + \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma_{[h,h']} \\
&= \sum_{h=1}^{H}\sigma_{[h,h]}\frac{N(n-H)-n}{NH^2} + \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} \\
&\leq \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} \leq 0,
\end{aligned}$$

where $\approx$ is used to indicate approximate equality since we replace limit f with its finite value for some n. This inequality yields that

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} \leq \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]} = \mathrm{var}\left(\sqrt{n}\hat{\mu}_0\right)$$

which completes the proof.

Proof of Theorem 2: We first look at the expected value of T_1 and T_2 under design-0 and design-2. Using conditional expectation, the expected value of T_1 reduces to

$$\begin{aligned}
E(T_1) &= \frac{1}{E\left(\frac{I_1 I_2}{d_n^2}\right)}\sum_{h=1}^{H}\sum_{h\neq h'}^{H} E\left(\frac{I_h I_{h'} M_h M_{h'}}{M_h M_{h'} d_n^2}\right) E\left(X_{[h]} - X_{[h']}\right)^2 \\
&= \begin{cases} 2(H-1)\sum_{h=1}^{H}\sigma^2_{[h]} + 2H\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 & \text{for design-0} \\ 2(H-1)\sum_{h=1}^{H}\sigma^2_{[h]} + 2H\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 - 2\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} & \text{for design-2.} \end{cases}
\end{aligned}$$

In a similar fashion, one can show that

$$\begin{aligned}
E(T_2) &= \sum_{h=1}^{H} E\left(\frac{HI_h^* M_h(M_h-1)}{M_h(Mh-1)d_n^*}\right) 2E\left\{\left(X_{[h]1} - \mu_{[h]}\right) - \left(\mu_{[h]} - X_{[h]2}\right)\right\}^2 \\
&= \begin{cases} 2HE\left(\frac{I_1^*}{d_n^*}\right)\sum_{h=1}^{H}\sigma^2_{[h]} & \text{for design-0} \\ 2HE\left(\frac{I_1^*}{d_n^*}\right)\left\{\sigma^2_{[h]} - \sum_{h}^{H}\sigma_{[h,h]}\right\} & \text{for design-2.} \end{cases}
\end{aligned}$$

Since I_h^*/d_n^*, $h = 1,\ldots,H$, are identically distributed, we have $E\left(I_1^*/d_n^*\right) = 1/H$. It follows that $E(T_2) = 2\sum_{h=1}^{H}\sigma^2_{[h]}$ for design-0 and $E(T_2) = 2\sum_{h=1}^{H}\sigma^2_{[h]} - 2\sum_{h}^{H}\sigma_{[h,h]}$ for design-2.

For the proof of $\bar{\sigma}^2$ we first observe that $E(T_1) + E(T_2) = 2H^2\sigma^2$ for design-0 and $E(T_1) + E(T_2) = 2NH^2\sigma^2/(N-1)$ for design-2. The proofs then follows from these equalities.

We complete the proof of the unbiasedness of $\hat{\sigma}^2_{\hat{\mu}_0}$ and $\hat{\sigma}^2_{\hat{\mu}_2}$ and $\bar{\sigma}^2_{\hat{\mu}_2}$ by inserting $E(T_1)$ and $E(T_2)$ in equations (2.5), (2.6) and (2.7).

The proofs for RSS estimators are similar. Hence, they are omitted.

References

Ahn, S., Lim, J. and Wang, X. (2014). The student's t approximation to distributions of pivotal statistics from ranked set samples. *Journal of Korean Statistical Society*, 43, 643-652.

Al-Saleh, M.F., and Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 198-209.

Dastbaravarde, A., Arghami, N.R. and Sarmad, M. (2016). Some theoretical results concerning non-parametric estimation by using a judgment poststratication sample. *Communications in Statistics - Theory and Methods*, 45, 8, 2181-2203.

Deshpande, J.V., Frey, J. and Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 25-40.

Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference*, 141, 3632-3639.

Frey, J., and Feeman, T.G. (2012). An improved mean estimator for judgment post-stratification. *Computational Statistics and Data Analysis*, 56, 418-426.

Frey, J., and Feeman, T.G. (2013). Variance estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 65, 551-569.

Frey, J., and Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 63, 769-789.

Gokpinar, F., and Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 89-95.

Jafari Jozani, M., and Johnson, B.C. (2011). Design based estimation for ranked set sampling in finite population. *Environmental and Ecological Statistics*, 18, 663-685.

Jafari Jozani, M., and Johnson, B.C. (2012). Randomized nomination sampling in finite populations. *Journal of Statistical Planning and Inference*, 142, 2103-2115.

MacEachern, S.N., Stasny, E.A. and Wolfe, D.A. (2004). Judgment post-stratification with imprecise ranking. *Biometrics*, 60, 207-215.

McIntyre, G.A. (1952). A method for unbiased selective sampling using ranked-sets. *Australian Journal of Agricultural Research*, 3, 385-390.

McIntyre, G.A. (2005). A method of unbiased selective sampling using ranked-sets. *The American Statistician*, 59, 230-232.

Ozdemir, Y.A., and Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 89-99.

Ozdemir, Y.A., and Gokpinar, F. (2008). A new formula for inclusion probablities in median ranked set sampling. *Communications in Statisitcs - Theory and Methods*, 37, 2022-2033.

Ozturk, O. (2013). Combining multi-obsever information in partially rank-ordered judgment post-stratifed and ranked set samples. *The Canadian Journal of Statistics*, 41, 304-324.

Ozturk, O. (2014a). Estimation of population mean and total in finite population setting using multiple auxiliary variables. *Journal of Agricultural, Biological and Environmental Statistics*, 19, 161-184.

Ozturk, O. (2014b). Statistical inference for population quantiles and variance in judgment post-stratified samples. *Computational Statistics and Data Analysis*, 77, 188-205.

Ozturk, O. (2015). Distribution free two-sample methods for judgment-post stratified data. *Statistica Sinica*, 25, 1691-1712.

Ozturk, O. (2016). Estimation of a finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological and Environmental Statistics*, 21, 1, 181-202.

Ozturk, O., and Jafari Jozani, M. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data anlysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. and Taillie, C. (1995). Finite population correction for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 621-636.

Presnell, B., and Bohn, L.L. (1999). U-Statistics and Imperfect Ranking in Ranked Set Sampling, 10, 111-126.

Stokes, S.L., Wang, X. and Chen, M. (2007). Judgment post stratification with multiple rankers. *Journal of Statistical Theory and Applications*, 6, 344-359.

Wang, X., Lim, J. and Stokes, S.L. (2008). A nonparametric mean estimator for judgment post-stratified data. *Biometrics*, 64, 355-363.

Wang, X., Wang, K. and Lim, J. (2012). Isotonized CDF estimation from judgment post-stratification data with empty strata. *Biometrics*, 68, 194-202.

Wang, X., Stokes, L., Lim, J. and Chen, M. (2006). Concomitant of multivariate order statistics with application to judgment poststratification. *Journal of American Statistical Association*, 101(476), 1693-1704.

Wolfe, D.A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *ISRN Probability and Statistics*, doi:10.5402/2012/568385.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 263-281
Statistics Canada, Catalogue No. 12-001-X

Adaptive rectangular sampling: An easy, incomplete, neighbourhood-free adaptive cluster sampling design

Bardia Panahbehagh[1]

Abstract

This paper introduces an incomplete adaptive cluster sampling design that is easy to implement, controls the sample size well, and does not need to follow the neighbourhood. In this design, an initial sample is first selected, using one of the conventional designs. If a cell satisfies a prespecified condition, a specified radius around the cell is sampled completely. The population mean is estimated using the π- estimator. If all the inclusion probabilities are known, then an unbiased π- estimator is available; if, depending on the situation, the inclusion probabilities are not known for some of the final sample units, then they are estimated. To estimate the inclusion probabilities, a biased estimator is constructed. However, the simulations show that if the sample size is large enough, the error of the inclusion probabilities is negligible, and the relative π- estimator is almost unbiased. This design rivals adaptive cluster sampling because it controls the final sample size and is easy to manage. It rivals adaptive two-stage sequential sampling because it considers the cluster form of the population and reduces the cost of moving across the area. Using real data on a bird population and simulations, the paper compares the design with adaptive two-stage sequential sampling. The simulations show that the design has significant efficiency in comparison with its rival.

Key Words: Adaptive cluster sampling; Adaptive two-stage sequential sampling; Primary and secondary sampling units; Inclusion probability.

1 Introduction

Adaptive cluster sampling (ACS) is an efficient design for rare and clustered populations (Thompson 1990; Thompson and Seber 1996). ACS was introduced for quadrat-based sampling, where the study area is usually partitioned into non-overlapping quadrats for sample selection. Depending on the situation, these are called “cells” or “secondary sampling units” (SSUs). In the first phase of the design, an initial sample is selected using one of the conventional designs, usually simple random sampling without replacement (SRSWOR). The term “conventional designs” (Thompson and Seber 1996) refers to designs in which the procedure for selecting the sample does not depend on any observation of the main variable, such as SRSWOR, stratified sampling and systematic sampling. If a rare event (a cell whose value is at least as large as the prespecified condition C) is found after the initial sample is obtained, then sampling continues in the neighbourhood of that location with the hope of observing more rare events. The process of searching the neighbourhood is continued until no more rare events are found. This design has been shown to be useful for estimating the parameters of highly clustered and rare populations (Smith, Brown and Lo 2004). However, ACS has some disadvantages, including the following two:

- The final sample size is random and uncontrollable.
- Neighbourhoods must be defined and followed. Following neighbourhoods in ACS means searching unit by unit and level by level around the initial rare events to find all the rare events

1. Bardia Panahbehagh, Department of Mathematics, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran. E-mail: panahbehagh@khu.ac.ir.

that are around them in different directions. The shape of the environment and the cluster may lead to confusion for the sampler.

To overcome the first problem, many designs, such as two-stage ACS and incomplete or restricted ACS (IACS), have been introduced.

Thompson (1991) introduced stratified ACS, and Salehi and Seber (1997) developed two-stage ACS. Two-stage ACS is designed to select a fixed number of primary sampling units (PSUs) by SRSWOR in the first stage, and then to select a fixed number of SSUs in each selected PSU, also by SRSWOR, in the second stage. The condition that is to be adapted and the neighbourhood are defined in terms of secondary units (rather than primary units). Salehi and Seber considered two schemes, depending on whether the clusters are allowed to overlap primary-unit boundaries or not, that would later be more desirable for controlling the final sample size. Some other related designs have also been introduced; as they are not related to the discussion in this paper, they are not mentioned.

Salehi and Smith (2005) made an essential change to two-stage ACS, known as two-stage sequential sampling, and Brown et al. (2008) introduced an adaptive version of two-stage sequential sampling (ATS). In ATS, the allocation of second-stage efforts among PSUs is based on preliminary information from the sampled PSUs. Additional survey efforts are directed to PSUs where the SSUs in the initial sample meet a prespecified criterion, or condition C (e.g., an individual from the rare population is present). More precisely, d times the number of units that satisfy condition C in the initial sample in a PSU is dedicated to the respective PSU as an additional sample using SRSWOR. Therefore, ATS could almost overcome the two problems, since, in this design, the final sample size is limited, and there is no need to define and follow the neighbourhood. But, ATS does not directly employ clustering of the population. This means that in ATS, the additional sample is a random sample of the whole respective PSU–ATS depends on the size, shape and location of the PSUs. Other developments in ATS are not essential (in other words, they have not changed the special aspects of ATS), so there is no need to mention them here.

IACS, Brown and Manly (1998) suggested a restricted version of ACS to control final sample size. They put a limit on the final sample size prior to sampling by selecting the initial sample sequentially. Chao and Thompson (1999) and Su and Quinn (2003) imposed a restriction on the number of neighbourhood levels beyond each unit that satisfies the condition in the initial fixed-size sample. All the neighbours of the units that satisfy the condition in the initial sample are in the first neighbourhood level. In turn, all the neighbours that are to be added based on the units in the first neighbourhood level are considered to be in the second neighbourhood level, and so on. In brief, a cluster, as defined in conventional ACS, is allowed in IACS to be truncated at a predetermined distance from the unit in the initial sample that intersects it. These authors introduced a biased estimator for the population mean.

Interestingly, Yang (2011) and Yang et al. (2011) introduced an adaptive plot design to overcome disadvantages of ACS in practice, and to use and define the neighbourhood, especially in tree-sampling surveys. They aimed to improve the practicability of ACS while maintaining some of its major characteristics. According to Yang et al. (2011), “the plot design is based on a conditional plot expansion: a larger plot (by a pre-defined plot size factor) is installed at a sample point instead of the smaller initial plot

if a pre-defined condition is fulfilled." Their target population was a tree population, and their aim was to estimate its density (the number of objects per hectare). Their design was not planned for quadrat-based sampling. In addition, they assumed that they could survey an additional area (after they finished sampling) to calculate the inclusion probability of the required tree, but this would be impossible or costly in other surveys.

Chao and Thompson (1999) introduced IACS for the first time. This design enables ACS to function without measuring all members of a cluster. They introduced the design in graph-theory form. Because it uses neighbourhoods like ACS, the design is complicated to manage, and the calculation of inclusion probabilities is also complicated.

In this paper, a manageable version of ACS, which has positive aspects of the designs discussed above, is proposed. Adaptive rectangular sampling (ARS) is a practical, efficient and easy-to-calculate adaptive design that is able to find rare events, does not need to follow the neighbourhood and controls the final sample size well.

ARS is introduced in Section 2 of the paper. Section 3 contains a real case study and some simulations, and Section 4 concludes the paper and provides a complete discussion of the advantages and disadvantages of the design.

2 Adaptive rectangular sampling

Suppose a total population of N units partitioned into M primary sampling units (PSUs), each containing N_h secondary sampling units (SSUs). Let $\{(h,j), h = 1,2,\ldots,M; j = 1,2,\ldots,N_h\}$ denote the j^{th} unit in the h^{th} primary unit, with the associated measurement or count y_{hj}. Then, $\tau_h = \sum_{j=1}^{N_h} y_{hj}$ is the total of the y values for the h^{th} PSU, and $\mu = 1/N \sum_{h=1}^{M} \tau_h$ is the population mean.

Adaptive rectangular sampling (ARS) can be performed in a two-stage procedure. The first stage of the ARS design consists of selecting a conventional random sample, s_0, of size m, with M PSUs.

The first phase of the second stage consists of selecting an initial conventional sample, s_{1h}, of size n_{1h} in the h^{th} PSU, where $h \in s_0$.

In the second phase, all the SSUs around those in s_{1h} that satisfy condition C with the radius R are adaptively added, where $R \in \{1,2,\ldots,M_d\}$ and M_d is the maximum diameter of each PSU. Here, the definition of radius is different from the conventional definition. "Radius," for a cell, is defined based on all cells around it. For example, $R = 1$ refers to the first-level nearest neighbourhood, which consists of the eight SSUs around the cell, and $R = 2$ refers to the nearest and the second-nearest neighbourhoods, which consist of all 24 SSUs around the cell (8 SSUs for $R = 1$, plus 16 SSUs added for $R = 2$). If the cell is in a corner or close to a border of the PSU, these numbers are reduced (see Figure 2.1). Therefore, in the h^{th} PSU, there is an additional sample, s_{2h}, of random size n_{2h}. Now, the final sample is $s_h = s_{1h} \cup s_{2h}$, of random size $n_h = n_{1h} + n_{2h}$, inside the h^{th} PSU. This procedure can be performed under either an overlapping scheme and a non-overlapping scheme.

An estimator for the population mean

The π- estimator (the Horvitz–Thompson estimator) is an estimator for the population mean that requires inclusion probabilities for all sampled units. If all the inclusion probabilities are available, the following are used to estimate the population mean:

$$\hat{\mu} = \frac{1}{N}\sum_{h\in s_0}\frac{\hat{\tau}_h}{\pi_h},$$

where π_h is the inclusion probability of the h^{th} PSU, and

$$\hat{\tau}_h = \sum_{j\in s_h}\frac{y_{hj}}{\pi_{hj}},$$

where π_{hj} is the inclusion probability of the SSU (h, j). To use the π- estimator, π_{hj} must be calculated for all $\{(h, j); h \in s_0, j \in s_h\}$.

In addition, the variance of the estimator is

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{h=1}^{M}\sum_{h'=1}^{M}\left(\frac{\pi_{hh'} - \pi_h\pi_{h'}}{\pi_h\pi_{h'}}\right)\tau_h\tau_h' + \sum_{h=1}^{M}\frac{\mathrm{Var}(\hat{\tau}_h)}{\pi_h}\right],$$

where $\pi_{hh'}$ is the joint inclusion probability of the h^{th} and h'^{th} PSUs. An unbiased estimator for the above is

$$\widehat{\mathrm{Var}}(\hat{\mu}) = \frac{1}{N^2}\left[\sum_{h\in s_0}\sum_{h'\in s_0}\left(\frac{\pi_{hh'} - \pi_h\pi_h'}{\pi_h\pi_h'}\right)\frac{\hat{\tau}_h\hat{\tau}_h'}{\pi_{hh'}} + \sum_{h\in s_0}\frac{\widehat{\mathrm{Var}}(\hat{\tau}_h)}{\pi_h}\right],$$

where $\widehat{\mathrm{Var}}(\hat{\tau}_h)$ is

$$\widehat{\mathrm{Var}}(\hat{\tau}_h) = \sum_{j\in s_{h1}}\sum_{j'\in s_{h1}}\left(\frac{\pi_{hjj'} - \pi_{hj}\pi_{hj'}}{\pi_{hj}\pi_{hj'}}\right)\frac{y_{hj}y_{hj'}}{\pi_{hjj'}},$$

where $\pi_{hjj'}$ is the joint inclusion probability of (h, j) and (h, j') in the h^{th} PSU.

It is easy to calculate the inclusion probabilities for the first stage, especially when simple random sampling without replacement (SRSWOR) is used. In this situation, it is easy to see that

$$\pi_h = \frac{m}{M}, \quad \pi_{hh'} = \frac{m(m-1)}{M(M-1)};\ h \neq h', \quad \pi_{hh} = \pi_h.$$

To calculate π_{hj}, it is necessary to know how many of the cells around cell (h, j) within radius R satisfy condition C, because selecting them as the initial sample leads to selecting the cells around them as the final sample. It is necessary to introduce some new notations. In ARS, with the radius R, B_{hj} represents the event of unit (h, j) being selected as the final sample, and A_{hj} represents the event of unit (h, j) satisfying condition C and being selected as the initial sample. In addition, s_{hj} represents all the units that satisfy condition C and that would be adaptively added to the sample if A_{hj} occurs, including unit (h, j),

with the size f_{hj}. The size f_{hj} is partitioned as $f_{hj} = f_{hj1} + f_{hj2}$, where the former indicates the number of cells in s_{hj} that are available in the final sample (s_h) and the latter is defined as $f_{hj2} = f_{hj} - f_{hj1}$. However, no information about its units is available. $F_{\cdot}$ is defined like $f_{\cdot}$, but for all units (those satisfying condition C and those not satisfying condition C).

In addition, let $f_{hjj'}$ be the size of $s_{hjj'} = s_{hj} \cup s_{hj'}$, and $f_{hjj'1}$, $f_{hjj'2}$, and $F_{\cdot}$ be defined the same.

The sample s_{hj}, with the size f_{hj}, contains all the cells that lead to the selection of unit (h, j), and $s_{hjj'}$, with the size $f_{hjj'}$, contains all the cells that lead to the selection of at least one of (h, j) and (h, j') as the final sample.

Theorem 1: In ARS, with the radius R, for the h^{th} PSU and using SRSWOR to select the initial sample of size n_{1h},

$$\pi_{hj} = 1 - \frac{\binom{N_h - f_{hj}}{n_{1h}}}{\binom{N_h}{n_{1h}}} \tag{2.1}$$

$$\pi_{hjj'} = 1 - \frac{\binom{N_h - f_{hj}}{n_{1h}}}{\binom{N_h}{n_{1h}}} - \frac{\binom{N_h - f_{hj'}}{n_{1h}}}{\binom{N_h}{n_{1h}}} + \frac{\binom{N_h - f_{hjj'}}{n_{1h}}}{\binom{N_h}{n_{1h}}}. \tag{2.2}$$

For the proof of the theorem, see the Appendix.

Here, only one problem arises: f_{hj} is known only in the initial sample that satisfies the condition. However, other samples (those that are adaptively added) have partial information about f_{hj} (i.e., f_{hj1}). Let $\mathrm{Bin}(n, p)$ stand for a binomial distribution based on the independent Bernoulli variable n with parameter p. To estimate f_{hj}, f_{hj2} represents the number of successes (the number of units that satisfy condition C) in F_{hj2} trials (by searching F_{hj2} units); the independency and identicality (iid) of the trials are assumed. The latter assumption (iid) is for simplifying the calculations. This, of course, leads to bias in estimating some of the inclusion probabilities and, so, to bias for the respective π- estimator. With all the above assumptions, F_{hj2} would be considered a random variable with a binomial distribution, as follows:

$$f_{hj2} \sim \mathrm{Bin}\left(F_{hj2}, p_{hj}\right),$$

where p_{hj} is the probability of satisfying condition C for all cells in the h^{th} PSU around the j^{th} cell with the radius R. Then,

$$E\left(f_{hj}\right) = f_{hj1} + E\left(f_{hj2}\right) = f_{hj1} + p_{hj}F_{hj2}.$$

Calculating π_{hj} leads to two situations:

- If the cell satisfies condition C and belongs to the initial sample, or is adaptively added and is located in a place in the final sample that contains complete information about the area around it, then it is possible to calculate the inclusion probability precisely from the information in the final sample.
- If the final sample does not contain enough information to calculate the inclusion probability, two strategies are proposed:
    - There is partial information about f_{hj}, this means that everything is known about F_{hj1}, and only F_{hj2} needs to be investigated. For F_{hj2}, only knowledge about how many of the cells satisfy the condition is required; there is no need for exact knowledge about y. For example, if condition C is defined as $y > 0$, it is necessary to know only how many cells of F_{hj2} are nonempty. If this is easy to investigate, the exact inclusion probabilities for all of the units in the sample can be calculated.
    - It is not possible to calculate the inclusion probabilities, π_{hj} can be estimated as (see equation 2.1)

$$\hat{\pi}_{hj} = 1 - \frac{\binom{N_h - E\left(f_{hj}\right)}{n_{1h}}}{\binom{N_h}{n_{1h}}} = 1 - \frac{\binom{N_h - \left(f_{hj1} + p_{hj}F_{hj2}\right)}{n_{1h}}}{\binom{N_h}{n_{1h}}}. \qquad (2.3)$$

Using p_{hj}, or, in other words, assuming different probabilities for different cells, leads to tedious calculations. Estimating p_{hj} can be done based on the spatial pattern of the population. For example, in the case of clustered populations, it may be reasonable to assume two kinds of p_{hj}, one for units in the sample satisfying condition C and one for units not satisfying condition C, so that greater probability is provided for the units satisfying condition C. A wide class of spatial patterns can be assumed in estimating p_{hj}, but, here, to have a simple and understandable strategy, $p_{hj} = p_h$ is assumed for all units in the h^{th} PSU. Therefore, p_h is the probability of satisfying condition C for units in the h^{th} PSU. It may be possible to guess p_h from previous information or to estimate it without bias from the initial sample as the portion of the units in the initial sample in the h^{th} PSU that satisfy condition C.

Estimating p_h based on the initial sample is a common procedure in adaptive designs (for example, see Brown et al. [2008]). For rare populations, such estimations might be imprecise. Practically, however, this is not a serious problem in ARS, because, for initial-sample units that satisfy the condition, it is possible to calculate inclusion probabilities without error (p_h is not required). Furthermore, p_h is not required in adaptively added units with $y = 0$. For some adaptively added units with $y > 0$, p_h has an insignificant role in calculating inclusion probabilities. The example in the next subsection and the simulation results in Section 3 confirm such assertions.

For $\pi_{hjj'}$ (as for π_{hj}), if, depending on the final sample, there is enough information to calculate f_{hj}, $f_{hj'}$ and $f_{hjj'}$, then it is enough to use equation 2.2. If there is partial information about $f_{hjj'}$, then

$$f_{hjj'2} \sim \text{Bin}\left(F_{hjj'2}, p_h\right),$$

and then

$$E\left(f_{hjj'}\right) = f_{hjj'1} + E\left(f_{hjj'2}\right) = f_{hjj'1} + p_h F_{hjj'2}.$$

And, it is enough to replace the respective "$f_{\cdot}$"s with "$E\left(f_{\cdot}\right)$"s in equation 2.2. Without the assumption of $p_{hj} = p_h$, $p_{hjj'}$ should perhaps be used instead of p_h for estimating $f_{hjj'}$. This would make the calculations more difficult. A reduction in precision (assuming constant p_h for all the units in the h^{th} PSU) allows such a simple sampling strategy to be presented.

Discussing an example can help clarify all of the above formulas and calculations.

Discussion of an example

For this example, see the top-left PSU in the right plot in Figure 2.1, where $N = 112$, $N_h = 28$, $h = 1$ and $n_{11} = 2$. Assume that it is necessary to calculate π_{hj} for two units in Figure 2.1, with $R = 1$. First, for the initial sample with $y = 6$, it is easy to see that $F_{hj} = 6$, that it has five cells around it plus itself, and that five of them satisfy the condition $\left(f_{hj} = 5\right)$. This information is available at the end of the sampling in the final sample. Therefore,

$$\pi_{y=6} = 1 - \frac{\binom{28-5}{2}}{\binom{28}{2}} \simeq 0.33.$$

For an adaptively added sample, like $y = 248$, as discussed earlier, there is partial information (see Figure 2.2). Here, $F_{hj} = 9$ (the blue cells in part B) and $F_{hj1} = 6$ (the orange cells in part C). From the final sample, $f_{hj1} = 5$ is also known (the positive response in the orange cells in part C). In addition, $F_{hj2} = 3$ (the blue cells in part C), but f_{hj2} is not known. To estimate it, p_1 would be estimated from the initial sample as $p_1 = 1/2 = 0.5$ (see the green cells in the first PSU in Figure 2.1), then

$$\hat{\pi}_{y=248} = 1 - \frac{\binom{28-(5+0.5\times 3)}{2}}{\binom{28}{2}} \simeq 0.42.$$

But, according to the population, the following can be calculated:

$$\pi_{y=248} = 1 - \frac{\binom{28-7}{2}}{\binom{28}{2}} \simeq 0.44,$$

which is very close to the estimation.

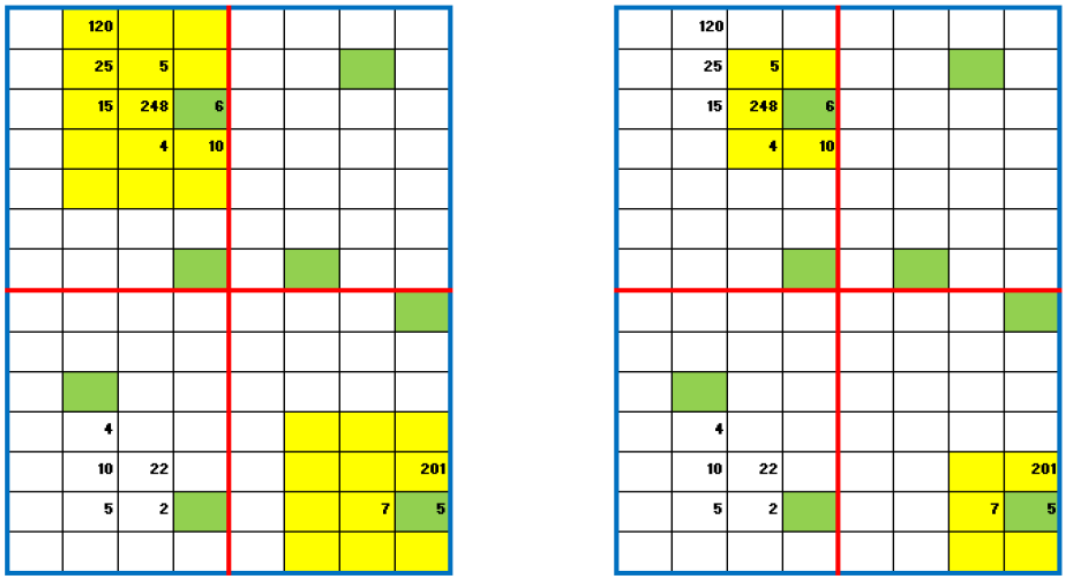

Figure 2.1 $N = 112$, $N_h = 28$, $M = 4$, $m = 4$ and $n_{1h} = 2$. **The green SSUs are the initial sample, and the yellow cells are the adaptively added sample. The right plot indicates non-overlapping ARS with** $R = 1$, **and the left plot indicates** $R = 2$, **where condition** C **is defined as** $y > 0$. **Numbers show respective** y **values for the cells.**

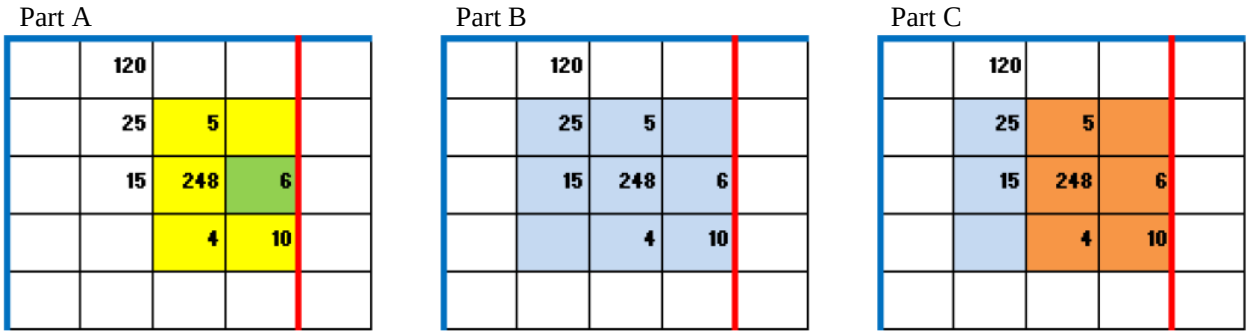

Figure 2.2 Inclusion probability for $y = 248$. **Part A indicates a part of the final sample. Part B (the blue cells) indicates the cells that must be known for the inclusion probability of** $y = 248$. **In part C, the orange cells are those that must be known and for which information is available, and the blue cells are those that must be known but for which no information is available.**

Now, assume that the goal is to calculate a joint probability, $\pi_{y=6,y'=5}$ (see Figure 2.3). Here, according to equation 2.2, $f_{hj} = 5$ (part B), $f_{hj'} = 3 + f_{hj'2}$, $F_{hj'2} = 5$ (the blue cells in part C), $f_{hjj'} = 5 + f_{hjj'2}$ and $F_{hjj'2} = 5$ (part D), and $E\left(f_{hjj'2}\right) = 5 \times 0.5 = 2.5$. From the information in the sample,

$$\hat{\pi}_{y=6,y'=5} = 1 - \frac{\binom{28-5}{2}}{\binom{28}{2}} - \frac{\binom{28-(3+0.5\times 5)}{2}}{\binom{28}{2}} + \frac{\binom{28-(5+0.5\times 5)}{2}}{\binom{28}{2}} \simeq 0.22.$$

With complete information about the population,

$$\pi_{y=6,y'=5} = 1 - \frac{\binom{28-5}{2}}{\binom{28}{2}} - \frac{\binom{28-6}{2}}{\binom{28}{2}} + \frac{\binom{28-8}{2}}{\binom{28}{2}} \simeq 0.22,$$

which shows no error to two decimal places. By writing a code for "$f_\cdot$"s and "$F_\cdot$"s, the π- estimator can be calculated easily.

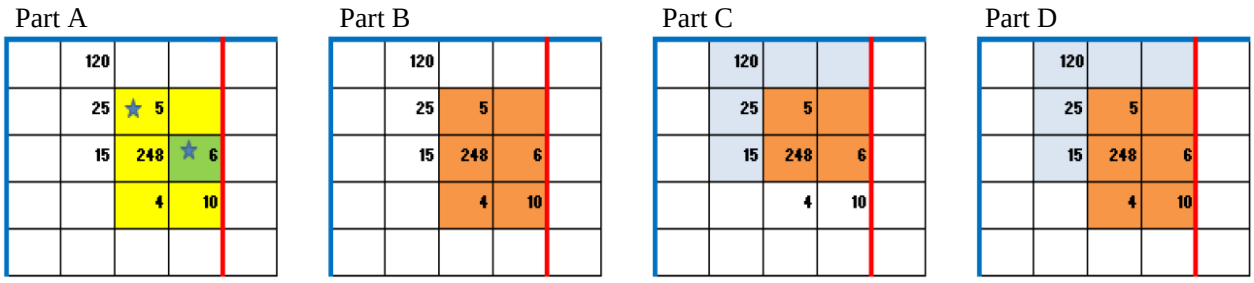

Figure 2.3 Joint inclusion probability for $y = 6$ and $y' = 5$. Coloured cells indicate the information that is required; orange cells indicate available information and blue cells indicate the information that needs to be estimated from the final sample.

3 A real case study and simulations

In this section, adaptive rectangular sampling (ARS) is evaluated and compared with adaptive two-stage sequential sampling (ATS) and two-stage simple random sampling without replacement (TSS). Here, ARS is not compared with adaptive cluster sampling (ACS) for two reasons: first, Salehi and Smith (2005) compared ATS with two-stage ACS, and, second, it is not fair to compare ARS with ACS or even incomplete ACS because ACS needs to define, use and follow the neighbourhood, while ARS does not.

If ATS is a design free of neighbourhood, then ARS satisfies this condition too, because, if a sampler can recognize the border of the cells, or, in other words, can distinguish secondary sampling units (SSUs), the sampler can also recognize an SSU with its radius area. In addition, the area to be surveyed may be specified before samples are taken. Based on a map of the area, it is possible to use SRSWOR for the SSUs

and the area around them if the SSUs satisfy the condition. Because the sampler need not return to the area to take the second phase of the sample (according to the ATS process), ARS seems to be easier and less costly than ATS. For a better comparison, the cost factor should be taken into consideration.

The comparison is done using two kinds of data: a real case study and simulation cases.

Here, efficiency is defined as

$$\text{eff}(.) = \frac{\text{MSE}(\bar{y}_{\text{tss}})}{\text{MSE}(.)},$$

where $\bar{y}_{\text{tss}}$ is the conventional mean estimator in TSS, MSE stands for mean square error and "." stands for one of the following:

- $\hat{\mu}_{\text{ATS}}$: Murthy's estimator in ATS, which is unbiased, and which will be referred to as "ATS". This estimator, for the mean of the h^{th} primary sampling unit (PSU), would be presented as

$$\hat{\mu}_{\text{ATS}} = \hat{q}_1 \bar{y}_{hc} + (1 - \hat{q}_1)\bar{y}_{hc'}, \tag{3.1}$$

  where $\hat{q}_1$ is the proportion of the units that satisfy condition C in the initial sample, and $\bar{y}_{hc}$ and $\bar{y}_{hc'}$ are the means of the final-sample units satisfying condition C and not satisfying condition C, respectively, in the h^{th} PSU. In this estimator, the first portion of the unit satisfying condition C is estimated from the initial sample, and the respective value is adapted to the mean of the sample satisfying condition C to construct the estimator.
- $\hat{\mu}_{\text{ARS}}$: the π- estimator in ARS, which is unbiased.
- $\hat{\mu}_{\text{ARS}.\hat{\pi}}$: the π- estimator in ARS, with π estimated from the final sample using equation 2.3. It is not unbiased, and its relative bias is defined as $\text{Rbias}.\hat{\mu}_{\text{ARS}.\hat{\pi}} = \left|\left(\mu_y - \hat{\mu}_{\text{ARS}.\hat{\pi}}\right)/\mu_y\right|$.

From now on, the acronym "ARS" refers to both $\hat{\mu}_{\text{ARS}}$ and $\hat{\mu}_{\text{ARS}.\hat{\pi}}$.

Furthermore, two formulas are used for the error in estimating inclusion probabilities in $\hat{\mu}_{\text{ARS}.\hat{\pi}}$:

- $\hat{\mu}_{D.Inclu}$: this shows the mean of the difference between real inclusion probabilities and the respective estimation (i.e., the mean of $e_{hj} = \pi_{hj} - \hat{\pi}_{hj}$ for all the sample units)
- $\hat{\mu}_{AD.Inclu}$: this shows the mean of the absolute difference between real inclusion probabilities and the respective estimation (i.e., the mean of $\left|e_{hj}\right| = \left|\pi_{hj} - \hat{\pi}_{hj}\right|$ for all the sample units).

A non-overlapping scheme is used in this section.

A real case study on a blue-winged teal population

Smith, Conroy and Brakhage (1995) used a population of blue-winged teal to evaluate ACS. The population comes from comprehensive counts, which were made from helicopters from December 13 to 15, 1992, in central Florida. The blue-winged teal population is extremely clustered, with a total of $N = 200$

units (Figure 3.1). A simulation study found ACS to be efficient for this population, in the sense that the variance of the estimator is smaller than in simple random sampling (Smith, Conroy and Brakhage 1995).

The population was partitioned into $M = 8, 4, 2$ PSUs. ARS, ATS and TSS were performed in the population with different values for m, n_{1h}, R and d (a multiple in ATS that indicates the size of the additional sample in the second phase for units satisfying condition C), with 25,000 simulations for each combination of values. For a fair comparison, d was chosen in such a way that the expected final sample sizes for ATS and ARS were almost the same. For TSS, the sample size in each simulation was the same as that for ARS. The expected sample sizes were calculated using Monte Carlo simulations. It is notable that in ARS, if two or more adaptively added samples overlapped, the overlap was measured once. Practically, if there is overlap in the sample, the relevant cells must be sampled and measured only once.

Results are presented in Tables 3.1 and 3.2. For information about the MSEs of the estimators, $\mathrm{MSE}(\bar{y}_{\mathrm{tss}})$ is presented in the results. With this MSE and the efficiency of the estimators, the MSEs of the other estimators are easy to calculate. The results are noteworthy: ARS was better than ATS in all situations. ARS, unlike ATS, was also always more efficient than TSS. The efficiency of ARS was sometimes seven or eight times that of TSS, whereas this number was at most around two and a half for ATS. For more than 55% of the cases, the efficiency of ARS was greater than 2, whereas this was true of less than 5% of the cases for ATS.

The relative bias of $\hat{\mu}_{\mathrm{ARS}.\hat{\pi}}$ is acceptable for most of the cases; it may be unacceptable for a few cases with a little sample size. For around 61% of the cases, the relative bias was less than 0.03, and, for around 92% of them, the relative bias was less than 0.07.

Efficiency improved by increasing the radius R, and a larger radius R was proper for larger PSUs. In this population, there are two important clusters at the top of the population plot. With $R = 2$, selecting one of the cells in a large PSU as the initial sample led to the selection of all of them. That is why $R = 2$, with a large enough initial sample size, showed such significant efficiency.

In addition, the number of PSUs in the first stage was important, and the results indicate that more PSUs lead to efficiency improvements. As discussed before, the efficiency of ATS depends on the size, shape and location of the PSUs. When the population could not be partitioned into some empty and full PSUs, ATS was not as efficient (see populations 2 and 4). But as ARS uses the cluster form of the population, it is not as dependent on PSUs and could even perform in a population with one PSU, which would be meaningless for ATS.

In addition, for Population 1, $\hat{\mu}_{AD.Inclu} = 0.025$ and $\hat{\mu}_{D.Inclu} = 0.002$; for Population 2, $\hat{\mu}_{AD.Inclu} = 0.023$ and $\hat{\mu}_{D.Inclu} = 0.005$; for Population 3, $\hat{\mu}_{AD.Inclu} = 0.024$ and $\hat{\mu}_{D.Inclu} = -0.004$; and, for Population 4, $\hat{\mu}_{AD.Inclu} = 0.025$ and $\hat{\mu}_{D.Inclu} = -0.008$. The mean of the inclusion probabilities for these simulations was around 0.22. According to $\hat{\mu}_{D.Inclu}$ and $\hat{\mu}_{AD.Inclu}$, the errors in estimating the inclusion probabilities seem to be almost negligible. This is why $\hat{\mu}_{\mathrm{ARS}.\hat{\pi}}$ was almost unbiased. The relative bias of $\hat{\mu}_{\mathrm{ARS}.\hat{\pi}}$ showed acceptable precision for "$\hat{\pi}_j$"s.

Table 3.1
Efficiency of the estimators, with $N = 200$, $C = 0$, $M = 8,4$

Population 1	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	Rbias.$\hat{\mu}_{ARS.\hat{\pi}}$	MSE$(\bar{y}_{tss})$
	1	4	1	6	7	1.32	2.05	1.62	0.13	81,076
			3	6	19	1.32	1.84	1.64	0.06	22,657
			5	5	30	1.25	1.79	1.69	0.03	12,865
			7	4	40	1.18	1.76	1.71	0.03	8,729
		8	1	6	13	1.22	1.92	1.51	0.14	35,171
			3	6	38	1.33	1.99	1.80	0.06	10,324
			5	5	59	1.35	2.13	2.04	0.04	5,502
			7	4	79	1.34	2.46	2.43	0.03	3,502
	2	4	1	15	10	2.43	3.02	3.04	0.02	71,670
			3	12	27	1.61	2.17	2.24	0.02	16,835
			5	9	40	1.33	1.99	2.05	0.00	9,219
			7	7	51	1.23	1.94	1.97	0.00	6,463
		8	1	15	20	2.37	3.00	3.01	0.02	32,290
			3	12	54	1.69	2.53	2.61	0.02	7,030
			5	9	80	1.52	2.98	3.05	0.02	3,570
			7	7	101	1.45	3.81	3.85	0.01	2,270
Population 2	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	Rbias.$\hat{\mu}_{ARS.\hat{\pi}}$	MSE$(\bar{y}_{tss})$
	1	2	4	6	13	1.06	1.90	1.58	0.11	36,208
			8	5	25	0.94	1.81	1.61	0.07	15,775
			10	5	30	1.03	1.82	1.66	0.05	12,434
			15	5	42	1.10	1.81	1.72	0.03	7,852
		4	4	6	26	1.00	1.92	1.60	0.11	16,911
			8	5	49	0.98	2.00	1.79	0.07	7,339
			10	5	60	1.00	2.05	1.89	0.06	5,475
			15	4	84	1.03	2.39	2.31	0.03	3,180
	2	2	4	12	20	1.29	2.51	2.52	0.00	28,341
			8	11	36	1.04	2.07	2.08	0.00	11,024
			10	10	42	1.03	2.02	2.02	0.01	8,521
			15	8	55	1.03	1.96	1.96	0.01	5,352
		4	4	12	40	1.04	2.32	2.32	0.02	11,725
			8	11	71	0.99	2.34	2.35	0.00	4,507
			10	10	84	1.01	2.59	2.60	0.00	3,338
			15	8	111	1.04	3.41	3.42	0.00	1,891

Table 3.2
Efficiency of the estimators, with $N = 200$, $C = 0$, $M = 4,2$

Population 3	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	Rbias.$\hat{\mu}_{ARS.\hat{\pi}}$	MSE$(\bar{y}_{tss})$
	1	2	4	6	14	1.11	1.73	1.40	0.18	34,451
			8	5	25	1.05	1.71	1.51	0.10	16,795
			10	5	31	1.09	1.69	1.56	0.08	13,671
			15	5	43	1.16	1.61	1.56	0.03	9,819
		4	4	6	27	1.26	2.18	1.73	0.18	15,357
			8	5	51	1.26	2.78	2.48	0.10	6,759
			10	5	62	1.34	3.18	2.94	0.08	5,078
			15	4	86	1.39	4.80	4.72	0.03	2,964
	2	2	4	12	21	1.30	1.90	1.95	0.05	26,117
			8	11	36	1.13	1.64	1.69	0.03	12,047
			10	10	43	1.11	1.56	1.60	0.02	9,872
			15	8	55	1.12	1.48	1.49	0.01	7,686
		4	4	12	42	1.43	2.55	2.64	0.05	10,818
			8	11	72	1.51	3.43	3.66	0.02	4,275
			10	10	85	1.55	4.29	4.57	0.01	3,183
			15	8	110	1.83	9.44	9.87	0.00	1,856
Population 4	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	Rbias.$\hat{\mu}_{ARS.\hat{\pi}}$	MSE$(\bar{y}_{tss})$
	1	1	10	6	17	0.93	1.68	1.32	0.13	27,053
			15	5	24	0.90	1.69	1.43	0.09	17,860
			20	5	31	0.91	1.63	1.45	0.08	13,802
			30	4	44	0.92	1.58	1.50	0.02	9,952
		2	10	6	34	0.98	2.16	1.69	0.12	11,851
			15	5	49	0.95	2.52	2.11	0.10	7,409
			20	5	62	0.97	2.91	2.58	0.06	5,153
			30	4	87	1.01	4.56	4.41	0.03	2,940
	2	1	10	14	27	1.01	1.70	1.69	0.02	18,908
			15	13	37	0.93	1.50	1.51	0.02	12,092
			20	10	45	0.84	1.42	1.43	0.01	9,338
			30	8	58	0.88	1.35	1.35	0.00	7,329
		2	10	14	53	1.08	2.38	2.41	0.02	7,512
			15	13	73	1.06	2.80	2.92	0.01	4,317
			20	11	90	1.03	3.59	3.79	0.00	2,933
			30	8	116	1.03	7.60	8.00	0.00	1,672

Artificial populations

The spatial pattern was generated with an R code following the Poisson cluster process (Brown 2003). The number of clusters was selected from a Poisson distribution, and cluster centres were randomly located throughout the site. Individuals within the cluster were located around the cluster centre at a random distance, following an exponential distribution, and in a random direction, following a uniform distribution. The parameters of the code were changed to generate three different populations. With the addition of the population in the example subsection of the paper, this subsection uses four artificial populations to evaluate ARS (see Figure 3.2):

- Population 5: rare and not clustered,
- Population 6: not rare, but clustered,
- Population 7: not rare and not clustered,
- Population 8: rare and clustered.

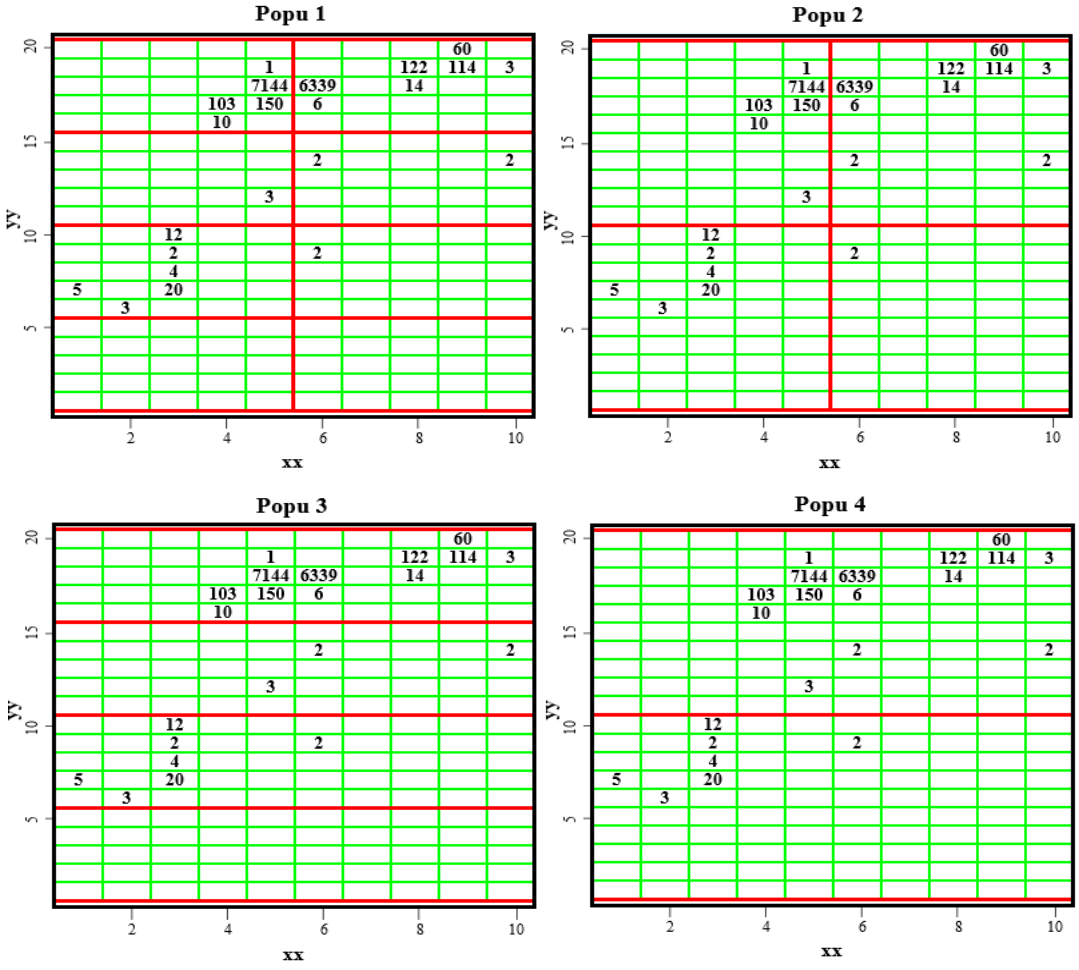

Figure 3.1 Bird populations. Numbers show the respective y values for the cells, with a mean of 70.61 and a variance of 453,709.52. Red lines indicate the borders of the PSUs.

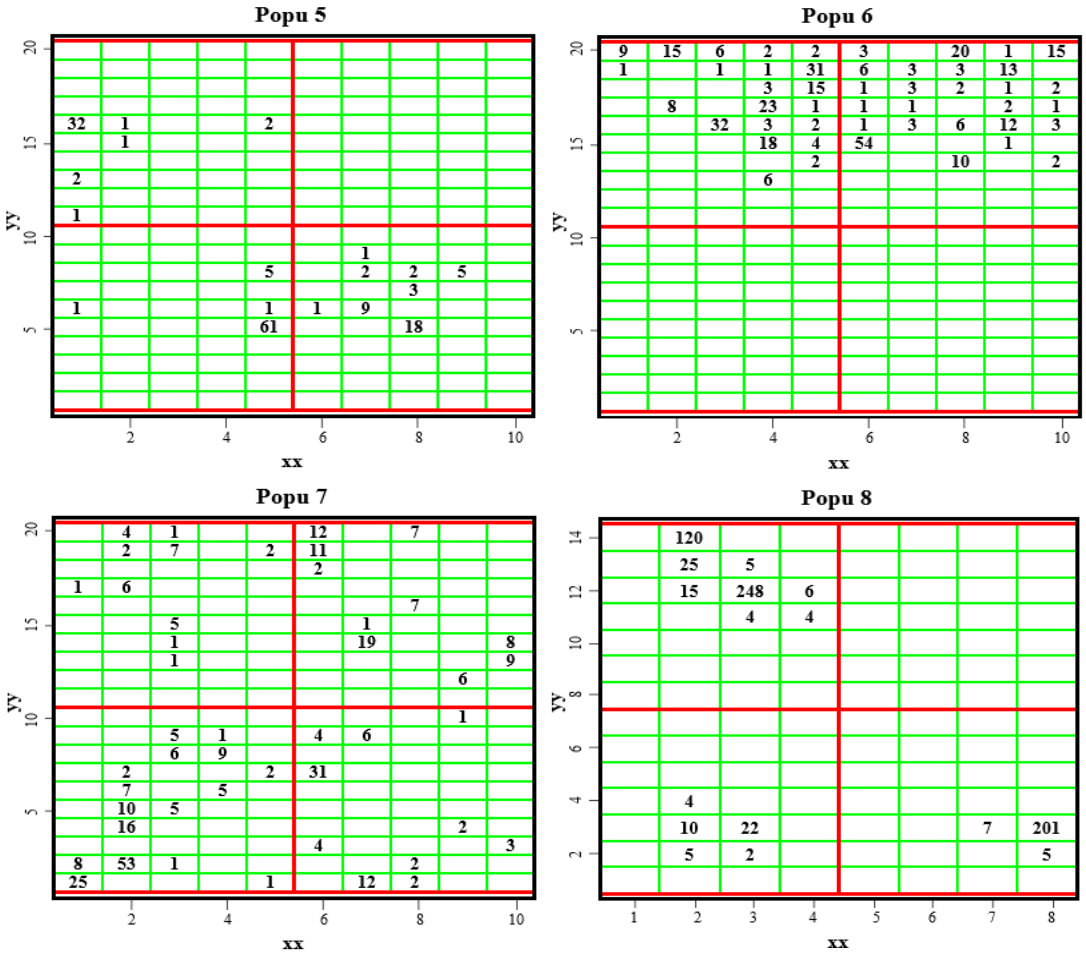

Figure 3.2 Artificial populations. Numbers show the respective y values for the cells, with 0.74, 1.76, 1.68 and 6.10 as the means, and 25.74, 35.34, 29.08 and 1,025.10 as the variances of the four populations, respectively. Red lines indicate the borders of the PSUs.

Results are presented in Tables 3.3 and 3.4. The relative bias of $\hat{\mu}_{\text{ARS}.\hat{\pi}}$ was acceptable for all cases.

Table 3.3
Efficiency of the estimators, with $N = 200$, $C = 0$, $M = 4$

Population 5	R	m	n_1	d	$E(n)$	$\hat{\mu}_{\text{ATS}}$	$\hat{\mu}_{\text{ARS}}$	$\hat{\mu}_{\text{ARS}.\hat{\pi}}$	Rbias.$\hat{\mu}_{\text{ARS}.\hat{\pi}}$	MSE$(\bar{y}_{\text{tss}})$
	1	2	4	6	12	0.94	1.53	1.59	0.01	2.29
			8	5	23	0.88	1.48	1.55	0.01	1.01
			10	5	28	0.88	1.45	1.52	0.00	0.77
			15	4	40	0.86	1.50	1.55	0.01	0.48
		4	4	6	24	0.88	1.44	1.50	0.01	1.01
			8	5	46	0.82	1.45	1.52	0.00	0.44
			10	5	57	0.80	1.48	1.54	0.01	0.34
			15	4	81	0.83	1.66	1.72	0.01	0.19
	2	2	4	14	18	1.01	1.30	1.38	0.04	1.93
			8	13	33	0.90	1.11	1.18	0.04	0.76
			10	11	39	0.82	1.07	1.14	0.02	0.56
			15	8	52	0.77	1.12	1.17	0.02	0.34
		4	4	14	37	0.80	1.01	1.08	0.03	0.74
			8	13	66	0.78	1.04	1.06	0.04	0.29
			10	10	78	0.67	1.02	1.03	0.03	0.22
			15	8	104	0.66	1.07	1.12	0.02	0.12

Table 3.3 (continued)
Efficiency of the estimators, with $N = 200$, $C = 0$, $M = 4$

Population 6	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	$\text{Rbias.}\hat{\mu}_{ARS.\hat{\pi}}$	$\text{MSE}(\bar{y}_{tss})$
	1	2	10	5	18	0.98	1.09	1.09	0.02	1.53
			15	4	31	0.96	1.08	1.12	0.01	1.35
			20	4	37	1.00	1.12	1.16	0.01	1.25
			30	3	49	1.02	1.13	1.16	0.01	1.14
		4	10	5	35	1.05	1.28	1.34	0.02	0.28
			15	4	62	1.09	1.56	1.73	0.00	0.17
			20	4	74	1.19	1.76	1.95	0.01	0.12
			30	2	97	1.25	2.48	2.68	0.01	0.06
	2	2	10	10	28	0.87	1.07	1.11	0.01	1.34
			15	7	42	0.90	1.15	1.18	0.01	1.24
			20	6	48	0.93	1.16	1.19	0.01	1.18
			30	5	57	0.99	1.15	1.16	0.00	1.12
		4	10	10	56	0.80	1.30	1.40	0.01	0.46
			15	7	85	0.91	2.76	2.97	0.01	0.23
			20	6	95	1.01	4.19	4.43	0.01	0.18
			30	4	114	1.18	12.19	12.43	0.00	0.12

Table 3.4
Efficiency of the estimators, with $N = 200,128$, $C = 0$, $M = 4$

Population 7	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	$\text{Rbias.}\hat{\mu}_{ARS.\hat{\pi}}$	$\text{MSE}(\bar{y}_{tss})$
	1	2	4	6	19	0.74	0.69	0.68	0.00	1.72
			8	6	34	0.84	0.76	0.75	0.02	0.84
			10	5	41	0.79	0.77	0.76	0.01	0.69
			15	3	56	0.73	0.85	0.84	0.01	0.50
		4	4	6	38	0.72	0.70	0.70	0.02	0.68
			8	5	69	0.66	0.57	0.57	0.01	0.28
			10	5	83	0.72	0.68	0.69	0.00	0.21
			15	4	112	0.69	0.76	0.75	0.02	0.11
	2	2	4	15	33	0.76	0.68	0.70	0.02	1.13
			8	10	55	0.57	0.62	0.64	0.00	0.54
			10	9	63	0.61	0.70	0.72	0.01	0.45
			15	8	79	0.63	0.81	0.82	0.02	0.36
		4	4	13	66	0.54	0.51	0.51	0.01	0.36
			8	10	110	0.40	0.45	0.48	0.02	0.13
			10	9	126	0.38	0.47	0.49	0.01	0.09
			15	8	156	0.31	0.51	0.53	0.01	0.04
Population 8	R	m	n_1	d	$E(n)$	$\hat{\mu}_{ATS}$	$\hat{\mu}_{ARS}$	$\hat{\mu}_{ARS.\hat{\pi}}$	$\text{Rbias.}\hat{\mu}_{ARS.\hat{\pi}}$	$\text{MSE}(\bar{y}_{tss})$
	1	2	3	6	11	0.99	1.97	1.85	0.05	95.39
			5	5	17	0.93	1.88	1.84	0.01	55.29
			10	4	30	0.96	1.72	1.73	0.00	27.73
			13	3	36	0.96	1.60	1.61	0.00	22.02
		4	3	6	22	0.92	2.21	2.10	0.04	39.74
			5	5	35	0.86	2.49	2.49	0.02	21.13
			10	3	59	0.85	4.28	4.36	0.00	8.12
			13	3	71	1.00	6.22	6.27	0.00	5.24
	2	2	3	13	17	1.11	1.68	1.74	0.02	78.55
			5	10	25	0.88	1.50	1.53	0.01	40.42
			10	7	37	0.87	1.41	1.41	0.01	21.05
			13	5	42	0.90	1.34	1.34	0.00	17.92
		4	3	13	34	0.97	1.70	1.75	0.03	27.92
			5	10	49	0.80	1.82	1.87	0.01	13.18
			10	7	74	0.80	3.35	3.39	0.00	4.86
			13	5	83	0.80	5.03	5.06	0.00	3.24

In Population 5, ATS was not efficient at all (except in one situation), but ARS performed well and was always more efficient than both ATS and TSS. ARS was better with $R = 1$, relative to $R = 2$, because the population was not clustered and a large radius could have wasted the sample. However, because of a large cluster in the lower-right PSU, ARS was efficient for $R = 2$.

In Population 6, which was highly clustered but not rare, ATS was not as efficient as TSS in almost half of the cases, especially when the sample size was not very large. The performance of ARS was very good. Because of the large size of the clusters, ARS was better with $R = 2$ than $R = 1$, and, in one case, it was

12 times as efficient as TSS, while this number was 1.18 for ATS. With $R = 2$, if a nonempty cell was selected as the initial sample, many of the other nonempty cells would also be selected. With a large enough initial sample size, almost all of them would be selected, providing almost complete information on the population and higher efficiency in comparison with other designs.

In Population 7, which was an almost-ordinary population, ARS and ATS were not efficient. In this case, the population was not clustered, and ARS wasted the sample searching around cells with a response satisfying condition C that were almost all empty. In such situations, ATS is more efficient than ARS (this happened in some cases), since ATS spreads the additional sample size equally across all the PSUs.

Finally, in Population 8, which was rare and completely clustered, ATS was almost not efficient at all. Again, ARS performed very well; it was sometimes six times as efficient as TSS and ATS.

In addition, for Population 5, $\hat{\mu}_{AD.Inclu} = 0.025$ and $\hat{\mu}_{D.Inclu} = -0.004$; for Population 6, $\hat{\mu}_{AD.Inclu} = 0.020$ and $\hat{\mu}_{D.Inclu} = -0.010$; for Population 7, $\hat{\mu}_{AD.Inclu} = 0.027$ and $\hat{\mu}_{D.Inclu} = 0.006$; and, for Population 8, $\hat{\mu}_{AD.Inclu} = 0.016$ and $\hat{\mu}_{D.Inclu} = 0.004$. The means of the inclusion probabilities for the four populations were, respectively, 0.21, 0.31, 0.24 and 0.34. Again, the errors in estimating the inclusion probabilities were almost negligible.

Lastly, $\hat{\mu}_{\mathrm{ARS}.\hat{\pi}}$ showed significant efficiency, even higher than $\hat{\mu}_{\mathrm{ARS}}$ (sometimes for a larger sample size). Since $\hat{\mu}_{\mathrm{ARS}}$ is unbiased, $\hat{\mu}_{\mathrm{ARS}}$ is preferred when there is enough information to calculate it. When information for calculating "π_j"s is lacking, $\hat{\mu}_{\mathrm{ARS}.\hat{\pi}}$ is a very good alternative for estimating the population mean with almost no bias.

Costs are not discussed, because this factor favours ARS, which is a cheaper design in comparison with ATS and TSS. Since ARS is more efficient than the other designs without considering costs, it is obvious that with costs factored in, the efficiency of ARS would be higher again. On the other hand, if the costs of much travelling under ATS and TSS are almost the same as the cost of searching more cells to find whether they satisfy condition C (not measuring them exactly) to calculate the unbiased π-estimator, then the comparison is fair.

4 Discussion

The adaptive rectangular sampling (ARS) design is an adaptive design that is easy to manage, saves on travel, is easy to calculate, is neighbourhood-free, and controls the final sample size.

The design is adaptive because the final sample size depends on observed values, and the design is able to find rare clustered events.

It is easy to manage because it is straightforward in determining the places that should be investigated for the additional sample. The design uses the intuitive behaviour of field biologists – once they find a rare event, they want to search in the immediate neighbourhood. It is even easier than adaptive cluster sampling (ACS). In addition, ARS can perform in both a two-stage form and a one-stage form. With this design, unlike with adaptive two-stage sequential sampling (ATS), there is no need to worry about the size, location and shape of primary sampling units. Unlike ACS, incomplete ACS (IACS) and ATS, it is possible for ARS to indicate the entire potential sample that the samplers need to select before they start sampling.

As for the travel-saving feature of ARS, there is no difference between adaptive designs such as ACS, ARS and ATS in the first phase of the second stage, for selecting the initial sample. But, in the second phase, ARS travels between cells generally less than ACS and IACS (with its edge units) and, especially, much less than ATS and two-stage sampling, with equal sample sizes. Because of this feature, ARS would be appropriate for costly travelling surveys of clustered populations, regardless of its efficiency.

ARS is easy to calculate, because the inclusion probabilities for the final sample size are easy to calculate, and this means that the π- estimator can be used instead of Murthy's estimator. Murthy's estimator, equation 3.1, is strongly dependent on the size of the initial sample (and on estimating q); as initial samples could be small in some situations, this is a weakness of Murthy's estimator. Therefore, one of the advantages of ARS as a sequential design is its avoidance of Murthy's estimator and its use of the π- estimator instead. In addition, calculating the π- estimator in IACS is not very easy, because it is a little complicated to estimate "π_j"s (see Chao and Thompson 1999). As discussed in Section 2.1, it is easy to calculate or estimate "π_j"s in ARS, compared with the method used by Chao and Thompson (1999).

The design is neighbourhood-free, in the sense that it does not follow the neighbourhood as in ACS and IACS; this would be complicated for the sampler after certain steps. ARS is an easy design for samplers, especially for difficult environments. ACS has not yet been used on a routine basis in field surveys for forest inventory and biodiversity monitoring, as there are also practical difficulties in field implementation (Yang et al. 2011). A design like ARS may be more appropriate in such environments.

The design controls the final sample size well with the choice of radius R. This paper presents an easy version of ARS. ARS can be performed in different ways (e.g., someone could plan a design to sample around a cell instead of investigating all the cells around it). This is a suggestion for future work on ARS.

To use ARS, it is important to know that the population units are separated in a cluster form; otherwise, the design would waste the sample units. This is one of the disadvantages of ARS. An advantage of this design is its expansion of the definition of clusters. Because the designer can change the radius R, a cluster in ARS consists of units that are around each other even at a distance, and there is no need for them to be adjacent.

Compared with other designs, ARS has some of the same advantages. Like ACS, it takes advantage of clustering to find rare events; like ATS, it does not need to follow the neighbourhood. And, like IACS, it controls the final sample size well.

ARS is a new design, and it should be evaluated on real populations to enable researchers to find out its abilities, advantages and disadvantages.

Appendix

For π_{hj}, it is easy to see that

$$\pi_{hj} = P\left(B_{hj}\right) = P\left(\cup_{k\in s_{hj}} A_{hk}\right) = 1 - P\left(\cap_{k\in s_{hj}} A'_{hk}\right) = 1 - \frac{\binom{N_h - f_{hj}}{n_{1h}}}{\binom{N_h}{n_{1h}}},$$

and, for $\pi_{hjj'}$, using the fundamental probability principle,

$$\begin{aligned}
\pi_{hjj'} &= P\left(B_{hj} \cap B_{hj'}\right) = 1 - P\left(B_{hj} \cap B_{hj'}\right)' = 1 - P\left(B'_{hj} \cup B'_{hj'}\right) \\
&= 1 - \left[P\left(B'_{hj}\right) + P\left(B'_{hj'}\right) - P\left(B'_{hj} \cap B'_{hj'}\right)\right] \\
&= 1 - \left[1 - P\left(\cup_{k \in s_{hj}} A_{hk}\right) + 1 - P\left(\cup_{k \in s_{hj'}} A_{hk}\right) - \left(1 - P\left(\cup_{k \in s_{hjj'}} A_{hk}\right)\right)\right] \\
&= 1 - \frac{\binom{N_h - f_{hj}}{n_{1h}}}{\binom{N_h}{n_{1h}}} - \frac{\binom{N_h - f_{hj'}}{n_{1h}}}{\binom{N_h}{n_{1h}}} + \frac{\binom{N_h - f_{hjj'}}{n_{1h}}}{\binom{N_h}{n_{1h}}}.
\end{aligned}$$

References

Brown, J.A. (2003). Designing an efficient adaptive cluster sample. *Environmental and Ecological Statistics*, 10, 95-105.

Brown, J.A., and Manly, B.J.F. (1998). Restricted adaptive cluster sampling. *Environmental and Ecological Statistics*, 5, 49-63.

Brown, J.A., Salehi, M.M., Moradi, M., Bell, G. and Smith, D.R. (2008). An adaptive two-stage sequential design for sampling rare and clustered populations. *Population Ecology*, 50, 239-245.

Chao, C.T., and Thompson, S.K. (1999). Incomplete adaptive cluster sampling designs. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 345-350.

Salehi, M.M., and Seber, G.A.F. (1997). Two-stage adaptive cluster sampling. *Biometrics*, 53, 959-970.

Salehi, M.M., and Smith, D.R. (2005). Two-stage sequential sampling: A neighborhood-free adaptive sampling procedure. *Journal of Agricultural, Biological, and Environmental Statistics*, 10, 84-103.

Smith, D.R., Brown, J.A. and Lo, N.C.H. (2004). Application of adaptive sampling to biological populations. In *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters*, (Ed., W.L. Thompson), Covelo, California: Island Press. 75-122.

Smith, D.R., Conroy, M.J. and Brakhage, D.H. (1995). Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics*, 51, 777-788.

Su, Z., and Quinn, T.J., II (2003). Estimator bias and efficiency for adaptive cluster sampling with order statistics and a stopping rule. *Environmental and Ecological Statistics*, 10, 17-41.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.

Thompson, S.K. (1991). Stratified adaptive cluster sampling. *Biometrika*, 78, 389-397.

Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*, New York: John Wiley & Sons, Inc.

Yang, H. (2011). *New Adaptive Plot Designs for Sampling Rare and Clustered Populations*. Ph.D. thesis at the Faculty of Forest Sciences and Forest Ecology of Georg-August-Universität Göttingen.

Yang, H., Kleinn, C., Fehrmann, L., Tang, S. and Magnussen, S. (2011). A new design for sampling with adaptive sample plots. *Environmental and Ecological Statistics*, 8, 223-237.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 283-295
Statistics Canada, Catalogue No. 12-001-X

Unequal probability inverse sampling

Yves Tillé[1]

Abstract

In an economic survey of a sample of enterprises, occupations are randomly selected from a list until a number r of occupations in a local unit has been identified. This is an inverse sampling problem for which we are proposing a few solutions. Simple designs with and without replacement are processed using negative binomial distributions and negative hypergeometric distributions. We also propose estimators for when the units are selected with unequal probabilities, with or without replacement.

Key Words: Location; Horvitz-Thompson estimator; Negative binomial; Negative hypergeometric; Inverse design; Inclusion probability; Wage.

1 Problem

The problem arose as part of a question on Statistics Canada's new Job Vacancy and Wage Survey (JVWS). The JVWS comprises a wage component and a job vacancy component. The wage component looks at average wages, minimum wages, maximum wages and starting wages for various occupations.

The objective is to provide wage statistics by economic regions (economic regions are subdivisions of provinces). In the first stage, a sample of 100,000 business locations (also known as local units of enterprises) are selected using a Poisson design stratified by industry and economic region.

For simplicity, the term "enterprise" will be used in the rest of the document instead of "location," keeping in mind that Statistics Canada defines a location as "a production unit located at a single geographical location at or from which economic activity is conducted and for which a minimum of employment data are available."

For purposes of managing response burden, it is not possible to identify every occupation in each enterprise. Therefore, proposing a list of occupations and asking whether the listed occupations exist in an enterprise has been considered. Occupations can then be randomly drawn from the list and proposed successively to the head of the enterprise until r occupations have been reached. Since the most common occupations are of specific interest, it is useful to consider cases in which occupations are selected with unequal probabilities from the list in proportion to their prevalence in the total population. Note that this method was not implemented for Statistics Canada's Job Vacancy and Wage Survey. The survey decided to present a list, of fixed length, of occupations to the surveyed enterprises. Nevertheless, the theoretical properties of the proposed method remain of interest.

"Inverse sampling" refers to a scheme in which units are selected successively until a predetermined number of units with a certain characteristic is obtained. Inverse sampling must not be confused with rejective sampling. In rejective sampling, a sample is selected according to a design, and the sample is rejected if it does not have the desired characteristic (e.g., a specific sample size or an average equal to that of the population). The selection of samples is repeated until a sample with the desired property is obtained.

1. Yves Tillé, Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland. E-mail: yves.tille@unine.ch.

Inverse sampling raises a certain number of theoretical questions. How can such a design be implemented with equal or unequal inclusion probabilities? What is the probability of inclusion of an occupation within each enterprise? How can a variable of interest be estimated using a sample consisting of a few enterprises and a few occupations within them? How can the number of occupations in the enterprise be estimated? More generally, how can this survey be implemented and how can estimation be done?

The key issue is the way in which the occupations are selected. They may be selected using a simple design with or without replacement, or with unequal probabilities. One option would be to select the units with unequal probabilities using the sequential Poisson sampling method proposed by Ohlsson (1998) or the Pareto sampling method proposed by Rosén (1997). The inverse sampling problem has already been discussed by Murthy (1957), Sampford (1962), Pathak (1964), Chikkagoudar (1966, 1969), and Salehi and Seber (2001). However, the parameter to be estimated here is unique, since estimates of average revenue among all enterprises having a specific occupation are desired. We also propose a new unequal-probability inverse design without replacement.

This article is organized as follows: In Section 2, the problem is stated and the notation is defined. The equal probability case with replacement is discussed in Section 3, and the equal probability case without replacement is discussed in Section 4. The unequal probability case with replacement is developed in Section 5. A new selection method for the unequal probability case without replacement is presented in Section 6. Finally, Section 7 contains a short discussion.

2 Formalization of the problem

The following notation is used:

- U : a population of N enterprises, i.e., $U = \{1,\ldots,i,\ldots,N\}$ (U may denote the population of enterprises in an economic region),
- L : the list of occupations,
- M : the number of occupations in the list, i.e., the size of L,
- F_i : the list of occupations in enterprise i, with $F_i \subset L$,
- D_i : the list of occupations absent from enterprise i, with $D_i \subset L$, $F_i \cup D_i = L$ and $D_i \cap F_i = \varnothing$,
- Mp_i : the number of occupations in enterprise i, i.e., the size of F_i,
- r : the number of distinct occupations to be obtained in each enterprise,
- X_i : the number of failures before the r occupations in enterprise i are obtained by selecting the occupations using a given design.

The main objective is to estimate the average wage for an occupation in the total population. Let y_{ik} be the average wage for occupation k in enterprise i, and let z_{ik} be the number of employees with occupation k in enterprise i. The objective is to estimate the average wage for occupation k given by

$$\bar{Y}_k = \frac{\sum_{i \in U \mid F_i \ni k} z_{ik} y_{ik}}{\sum_{i \in U \mid F_i \ni k} z_{ik}}.$$

Assume that a sample of enterprises S_1 is selected from U using some given design with inclusion probabilities π_{1i}. In enterprise i, a sample of occupations S_i is selected using one of the designs described above with inclusion probability $\pi_{k|i}$. If the design is with replacement, $\pi_{k|i}$ represents the expected number of times that occupation k is selected in enterprise i.

$\bar{Y}_k$ can be estimated using a "ratio" type estimator (Hájek 1971):

$$\hat{\bar{Y}}_k = \frac{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik} y_{ik}}{\pi_{1i} \pi_{k|i}}}{\sum_{i \in S_1 \mid (S_i \cap F_i) \ni k} \frac{z_{ik}}{\pi_{1i} \pi_{k|i}}}.$$

Therefore, the probability that an occupation will be selected in an enterprise must be known. However, with an inverse type design, the probability is unknown and must therefore be estimated in order to estimate $\bar{Y}_k$. Since the inclusion probabilities appear in the denominator, it is preferable to estimate the inverses of $\pi_{k|i}$. In an enterprise, an occupation's probability of being selected decreases as the number of occupations increases. In addition, the probability depends on the inverse sampling design used in each enterprise.

3 Simple random sampling with replacement

Assume that enterprise i has proportion p_i of the occupations in the list in the enterprise. If the sample of occupations is drawn with replacement in enterprise i until r occupations in the enterprise have been identified, then X_i has a negative binomial distribution denoted by $X_i \sim NB(r, p_i)$. In that case,

$$\Pr(X_i = x_i) = \binom{r + x_i - 1}{x_i} p_i^r (1 - p_i)^{x_i},$$

with $x_i \in \mathbb{N} = \{0,1,2,3,\ldots\}, p_i \in [0,1], r \in \mathbb{N}^* = \{1,2,3,\ldots\}$. Furthermore,

$$\mathrm{E}(X_i) = \frac{r(1-p_i)}{p_i} \quad \text{and} \quad \mathrm{var}(X_i) = \frac{r(1-p_i)}{p_i^2}.$$

Let $A_{ik}, k \in L,$ be the number of times that unit k is selected in the sample taken from enterprise i. In a simple design with replacement of size $n,$ the values of A_{ik} have a multinomial distribution. Therefore,

$$\Pr(A_{ik} = a_{ik}, k \in L) = \frac{n!}{M^n} \prod_{k \in L} \frac{1}{a_{ik}!},$$

where $A_{ik} = 0, \ldots, n,$ and

$$\sum_{k\in L} a_{ik} = n.$$

If this multinomial vector is conditioned on a fixed size in a given part of the population, then

$$\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k\in F_i} A_{ik} = r\right) = \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k\in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k\in F_i} A_{ik} = r\right)}$$

$$= \frac{\dfrac{n!(1-p_i)^{(n-r)}}{(n-r)!M^r}\displaystyle\prod_{k\in F_i}\frac{1}{a_{ik}!}}{\dfrac{n!p_i^r(1-p_i)^{n-r}}{r!(n-r)!}}$$

$$= r!\left(\frac{1}{Mp_i}\right)^r \prod_{k\in F_i}\frac{1}{a_{ik}!},$$

with

$$\sum_{k\in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, the distribution remains multinomial and conditionally there is still a simple design with replacement.

With the procedure in which we draw with replacement until we obtain r occupations in enterprise i, we have

$$\mathrm{E}\left(A_{ik} \,|\, X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{if } k \in F_i \\ \dfrac{X_i}{M - Mp_i} & \text{if } k \in D_i. \end{cases}$$

In fact, conditionally on X_i, in F_i of size Mp_i, r occupations are selected and, in D_i of size $M(1-p_i)$, X_i occupations are selected.

In the case with replacement, what is calculated is not really an inclusion probability, but rather the expected value of A_{ik} which is denoted as $\pi_{k|i}$,

$$\pi_{k|i} = \mathrm{EE}\left(A_{ik} \,|\, X_i\right) = \frac{r}{Mp_i},$$

$k \in L$. The problem is that we know M, r and X_i, but not p_i. We can estimate p_i using the method of moments by solving $\mathrm{E}(X_i) = X_i$, which yields

$$X_i = \frac{r(1-\hat{p}_i)}{\hat{p}_i}$$

and therefore

$$\hat{p}_{i1} = \frac{r}{X_i + r}.$$

The maximum likelihood method provides the same estimator as the method of moments, but this estimator is biased (Mikulski and Smith 1976; Johnson, Kemp and Kotz 2005, page 222). If $r \geq 2$, the unbiased minimum variance estimator of p_i is

$$\hat{p}_{i2} = \frac{r-1}{X_i + r - 1}.$$

However, $1/\hat{p}_{i1}$ is unbiased for $1/p_i$.

Since we are using weights that are inverses of $\pi_{k|i}$, the inverses of $\pi_{k|i}$ are thus estimated as follows:

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} = \dfrac{M(r-1)}{r(X_i + r - 1)} & \text{if } k \in F_i \\ \dfrac{M(1-\hat{p}_{i2})}{X_i} = \dfrac{M}{X_i + r - 1} & \text{if } k \in D_i. \end{cases}$$

However, the case with replacement is not very satisfactory, because selecting r occupations with replacement does not necessarily result in r distinct occupations, since the same occupation may be selected more than once. Furthermore, sampling may be especially long if Mp_i is small. Therefore, sampling without replacement is preferred.

4 Simple random sampling without replacement

For the case without replacement, the notation used is the same as for the draw with replacement. The number of failures X_i therefore has a negative hypergeometric distribution. This probability distribution is little known, to the point that it has been presented as a "forgotten" distribution by Miller and Fridell (2007). This distribution is the counterpart to the negative binomial for the draw without replacement. The general framework is as follows: We consider a population of size M in which there are Mp_i favourable units, namely the occupations in the list that exist in the enterprise. If the draws are equal probability without replacement until r favorable units appear, then the negative hypergeometric variable, $X_i \sim NH(M, r, Mp_i)$, counts the number of failures before r favourable events occur.

The probability distribution is

$$\Pr(X_i = x) = p(x; M, r, Mp_i) = \frac{\binom{x+r-1}{x}\binom{M-x-r}{Mp_i - r}}{\binom{M}{Mp_i}},$$

where $x \in \{0,\ldots,M(1-p_i)\}$, $M \in \{1,2,\ldots\}$, $Mp_i \in \{1,2,\ldots,M\}$, and $r \in \{1,2,\ldots,Mp_i\}$.

$$\mathrm{E}(X_i) = \frac{Mr(1-p_i)}{Mp_i+1}, \mathrm{var}(X_i) = \frac{rM(1-p_i)(M+1)(Mp_i-r+1)}{(Mp_i+1)^2(Mp_i+2)}.$$

Again, A_{ik} denotes the number of times that unit k is selected in the sample. Now, the value of A_{ik} can be only 0 or 1. If n units are selected using a simple design without replacement in L, the sample design is defined as

$$\Pr(A_{ik} = a_{ik}, k \in L) = \binom{M}{n}^{-1},$$

where $a_{ik} \in \{0,1\}$, and

$$\sum_{k\in L} a_{ik} = n.$$

If the vector of A_{ik} is conditioned on a fixed size in one part of the population, we have

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k\in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k\in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k\in F_i} A_{ik} = r\right)} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \sum_{\substack{k\in D_i \\ \sum_{k\in F_i} A_{ik} = n-r \\ A_{ik}\in\{0,1\}}} \frac{1}{\binom{M}{n}} \\
&= \left[\frac{\binom{Mp_i}{r}\binom{M-Mp_i}{n-r}}{\binom{M}{n}}\right]^{-1} \frac{\binom{M-Mp_i}{n-r}}{\binom{M}{n}} \\
&= \binom{Mp_i}{r}^{-1},
\end{aligned}$$

with

$$\sum_{k\in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, we still have a simple design without replacement. In the procedure in which we draw without replacement until we obtain r occupations in enterprise i, we therefore have

$$\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{if} \quad k \in F_i \\ \dfrac{X_i}{M - Mp_i} & \text{if} \quad k \in D_i. \end{cases}$$

The inclusion probability is therefore

$$\pi_{k|i} = \mathrm{E}\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{r}{Mp_i} & \text{if} \quad k \in F_i \\ \dfrac{\mathrm{E}\left(X_i\right)}{M - Mp_i} = \dfrac{r}{Mp_i + 1} & \text{if} \quad k \in D_i, \end{cases}$$

for all $k \in L$. Again, the problem is that we know M, r and X_i, but not p_i. We can estimate p_i using the maximum likelihood method, through a numerical method.

Using the method of moments, an estimate can be obtained by solving for p_i in the equation $X_i = \mathrm{E}\left(X_i\right)$, that is,

$$X_i = \frac{Mr\left(1 - \hat{p}_i\right)}{M\hat{p}_i + 1}.$$

Hence

$$\hat{p}_{i1} = \frac{Mr - X_i}{M\left(r + X_i\right)}.$$

However, in a few lines it is verified that, if $r \geq 2$,

$$\hat{p}_{i2} = \frac{r - 1}{r + X_i - 1}$$

is unbiased for p_i.

Again, since we are using weights that are inverses of $\pi_{k|i}$. The inverses of the inclusion probabilities are thus estimated as follows:

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{M\hat{p}_{i2}}{r} = \dfrac{M\left(r-1\right)}{r\left(X_i + r - 1\right)} & \text{if} \quad k \in F_i \\ \dfrac{M\left(1 - \hat{p}_{i2}\right)}{X_i} = \dfrac{M}{X_i + r - 1} & \text{if} \quad k \in D_i. \end{cases}$$

These weights are also used in the estimator by Murthy (1957), which is unbiased (see also Salehi and Seber 2001). If $Mp_i < r$, all occupations will be selected in enterprise i and the estimated inclusion probabilities are then equal to 1.

5 Unequal probability sampling with replacement

Unequal probability sampling is not really more difficult to process when the draw is with replacement. Now let p_{ik} denote the probability of an occupation being drawn in each draw with

$$\sum_{k\in L} p_{ik} = 1.$$

Let P_i be the sum of p_{ik} limited to the occupations in enterprise i:

$$P_i = \sum_{k\in F_i} p_{ik}.$$

In this case, X_i has a negative binomial distribution with parameters r and P_i. Therefore,

$$\mathrm{E}(X_i) = \frac{r(1-P_i)}{P_i} \quad \text{and} \quad \mathrm{var}(X_i) = \frac{r(1-P_i)}{P_i}.$$

Let $A_{ik}, k \in L$ be the number of times that unit k is selected in the sample. In an unequal probability design with replacement of size n, the values of A_{ik} have a multinomial distribution. Therefore,

$$\Pr(A_{ik} = a_{ik}, k \in L) = n! \prod_{k\in L} \frac{p_{ik}^{a_{ik}}}{a_{ik}!},$$

where $A_{ik} = 0, \ldots, n,$ and

$$\sum_{k\in L} a_{ik} = n.$$

If this multinomial vector is conditioned on a fixed size in one part of the population, then

$$\begin{aligned}
\Pr\left(A_{ik} = a_{ik}, k \in F_i \,\middle|\, \sum_{k\in F_i} A_{ik} = r\right) &= \frac{\Pr\left(A_{ik} = a_{ik}, k \in F_i \text{ and } \sum_{k\in F_i} A_{ik} = r\right)}{\Pr\left(\sum_{k\in F_i} A_{ik} = r\right)} \\
&= \frac{\dfrac{n!(1-P_i)^{(n-r)}}{(n-r)!} \displaystyle\prod_{k\in F_i} \frac{p_{ik}^{a_{ik}}}{a_{ik}!}}{\dfrac{n! P_i^r (1-P_i)^{n-r}}{r!(n-r)!}} \\
&= r! \prod_{k\in F_i} \left(\frac{p_{ik}}{P_i}\right)^{a_{ik}} \frac{1}{a_{ik}!},
\end{aligned}$$

with

$$\sum_{k\in F_i} a_{ik} = r.$$

This shows that, if the sum of A_{ik} is conditioned on one part of the population, the distribution remains multinomial and conditionally there is still an unequal probability design with replacement.

With the procedure in which we draw with replacement until we obtain r occupations in enterprise i, we have

$$\mathrm{E}\left(A_{ik} \mid X_i\right) = \begin{cases} \dfrac{rp_{ik}}{P_i} & \text{if} \quad k \in F_i \\ \dfrac{X_i p_{ik}}{1-P_i} & \text{if} \quad k \in D_i. \end{cases}$$

The expected value of A_{ik} is

$$\pi_{k|i} = \mathrm{EE}\left(A_{ik} \mid X_i\right) = \frac{rp_{ik}}{P_i},$$

$k \in L$. The problem is that we know p_{ik}, r and X_i, but not P_i. We can estimate P_i using the method of moments by solving $\mathrm{E}\left(X_i\right) = X_i$, which gives

$$X_i = \frac{r\left(1-\hat{P}_i\right)}{\hat{P}_i}$$

and therefore

$$\hat{P}_{i1} = \frac{r}{X_i + r}.$$

The maximum likelihood method provides the same estimator as the method of moments, but this estimator is biased (Mikulski and Smith 1976; Johnson et al. 2005, page 222). In fact, the unbiased minimum variance estimator is

$$\hat{P}_{i2} = \frac{r-1}{X_i + r - 1}.$$

However, $1/\hat{P}_{i1}$ is unbiased for P_i.

Again, since we are using weights that are inverses of $\pi_{k|i}$. The inverses of $\pi_{k|i}$ are thus estimated as follows:

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{\hat{P}_{i2}}{rp_{ik}} = \dfrac{r-1}{\left(X_i + r - 1\right)rp_{ik}} & \text{if} \quad k \in F_i \\ \dfrac{1-\hat{P}_{i2}}{X_i p_{ik}} = \dfrac{1}{\left(X_i + r - 1\right)p_{ik}} & \text{if} \quad k \in D_i. \end{cases} \tag{5.1}$$

6 Unequal probability sampling without replacement

6.1 Sequential sampling without replacement

For the draw without replacement, the first problem is determining the design. One option is to use the method by Ohlsson (1995) called sequential Poisson sampling. This method involves generating M uniform random variables in the interval $[0,1]$, denoted u_{ik}. Next, we select the n units corresponding to the smallest values of $u_{ik}/\pi_{k|i}$. This method has the advantage of being usable for any sample size and providing a sequence of samples that are included in each other. Unfortunately, it only satisfies approximately the fixed inclusion probabilities. However, the approximations are very accurate according to the simulations given in Ohlsson (1995).

Methods have also been proposed by Sampford (1962) and Pathak (1964). We propose an exact solution to the problem in the sense that the inclusion probabilities are exactly satisfied. We begin by calculating the inclusion probabilities for a design of fixed size n with inclusion probabilities proportional to a strictly positive auxiliary variable $b_k, k \in L.$ The probabilities are determined by

$$\pi_{k|i}(n) = \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right),$$

where C_n is determined such that

$$\sum_{k \in L} \pi_{k|i}(n) = \sum_{k \in L} \min\left(1, C_n \frac{b_k}{\sum_{\ell \in L} b_\ell}\right) = n.$$

A simple algorithm for calculating these probabilities is described in Tillé (2006, page 19), among others. The probabilities can be calculated simply using the function `inclusionprobabilities` in the R sampling package.

A sequential selection method must therefore select a sample of size n with inclusion probabilities $\pi_{k|i}(n)$. It must then make it possible to go from size n to size $n+1$ by simply selecting an additional unit such that the completed sample has an inclusion probability of $\pi_{k|i}(n+1)$. It appears that the only method that allows that to be achieved is the elimination method (Tillé 1996). This method starts with the entire population (the list of occupations) and eliminates one unit in each step. In step $j = 1,\ldots,N,$ the unit is eliminated from among the remaining units with the probability

$$1 - \frac{\pi_{k|i}(N-j)}{\pi_{k|i}(N-j+1)}.$$

This method can thus be used to create a sequence of samples included in each other that verify the inclusion probabilities in relation to their size.

Therefore, we can simply apply the elimination method for sample size $n = 1$ so that the algorithm successively eliminates all the units. Taking them in the reverse order of elimination, we obtain a sequence of units. The first n units of the sequence are selected with inclusion probability $\pi_{k|i}(n)$. The appendix

contains a function written in R that can be used to generate this sequence. The code is executed in a simulation that shows that the probabilities obtained through simulations by applying this function are equal to the fixed inclusion probabilities for all sample sizes.

6.2 Inverse or negative design with unequal probabilities

Now that the design is defined, the inverse design can be defined. The units in the list of occupations are taken using the elimination method until r occupations in the enterprise are selected. In this case, the probability distribution of the number of failures X_i seems impossible to calculate. Calculating the conditional inclusion probability $\mathrm{E}\left(A_{ik} \mid X_i\right)$ is also problematic.

However, we can proceed by analogy and estimate the inclusion probabilities on the basis of expression (5.1) developed for the case with replacement, where p_{ik} can simply be replaced by

$$\frac{\pi_{k|i}\left(r+X_i\right)}{r+X_i}.$$

Therefore, we obtain

$$\widehat{1/\pi_{k|i}} = \begin{cases} \dfrac{(r-1)\left(r+X_i\right)}{r\left(X_i+r-1\right)\pi_{k|i}\left(r+X_i\right)} & \text{if} \quad k \in F_i \\ \dfrac{r+X_i}{\left(X_i+r-1\right)\pi_{k|i}\left(r+X_i\right)} & \text{if} \quad k \in D_i. \end{cases}$$

7 Discussion

The selection problem can therefore be resolved for all cases, with or without replacement and with equal or unequal probabilities. The proposed solution based on the elimination method respects the inclusion probabilities exactly, which is not true for Ohlsson's sequential sampling. The implementation is especially simple, since the program provides an ordered sequence of occupations to propose until the objective has been met.

The estimation issue is slightly more difficult. For the unequal probability sampling without replacement, we must make do with a heuristic solution. As well, it can be seen that, in the second stage, there tends to be lower inclusion probabilities in enterprises that have many occupations. This should lead us to select with greater probabilities the enterprises that may have a larger number of occupations, to avoid selecting occupations with probabilities that are too unequal.

Acknowledgements

The author wishes to thank Pierre Lavallée for submitting this interesting problem and providing thoughtful comments on an earlier version of this article. The author also thanks Audrey-Anne Vallée for her meticulous proofreading, and a referee and writer of *Survey Methodology* for their pertinent remarks, which made it possible to improve this article.

Appendix

```
#
# Load sampling package, which contains the function inclusionprobabilities().
#
library(sampling)
#
# The function returns a vector with the sequence numbers of the eliminations.
# The last (resp. first) unit eliminated is the first (resp. last)
# component of the vector.
# The function therefore provides the numbers of the units to be presented
# successively for the inverse selection.
# The argument x is the vector of values of the auxiliary variable used to calculate
# the inclusion probabilities.
#
elimination<-function(x)
       {
       pikb=x/sum(x)
       M = length(pikb)
       n = sum(pikb)
       sb = rep(1, M)
       b = rep(1, M)
       res=rep(0, M)
       for (i in 1:(M)) {
              a = inclusionprobabilities(pikb, M - i)
              v = 1 - a/b
              b = a
              p = v * sb
              p = cumsum(p)
              u = runif(1)
              for (j in 1:length(p)) if (u < p[j])
                     break
              sb[j] = 0
              res[i]=j
              }
       res[M:1]
       }
#
# 500,000 simulations with a size in a list of size M=20.
# By taking the first m components of vector v, we obtain a sample
# of size m.
#
M=20
x=runif(M)
Pik=array(0,c(M,M))
#
# Calculate the inclusion probabilities for all sample sizes from 1 to 20.
#
for(i in 1:M) Pik[i,]=inclusionprobabilities(x, i)
rowSums(Pik)

SIM=50000
SS=array(0,c(M,M))
for(i in 1:SIM)
{
S=array(0,c(M,M))
v=elimination(x)
for(i in 1:M) S[i,v[1:i]]=1
SS=SS+S
}
SS=SS/SIM
#
# Compare actual and empirical inclusion probabilities.
#
Pik
SS
SS-Pik
```

References

Chikkagoudar, M.S. (1966). A note on inverse sampling with equal probabilities. *Sankhyā,* A28, 93-96.

Chikkagoudar, M.S. (1969). Inverse sampling without replacement. *Australian Journal of Satistic*, 11, 155-165.

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, part on by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), page 326, Toronto, Canada. Holt, Rinehart, Winston.

Johnson, N.L., Kemp, A.W. and Kotz, S. (2005). *Univariate Discrete Distributions*. New York: John Wiley & Sons, Inc.

Mikulski, P.W., and Smith, P.J. (1976). A variance bound for unbiased estimation in inverse sampling. *Biometrika*, 63(1), 216-217.

Miller, G.K., and Fridell, S.L. (2007). A forgotten discrete distribution? Reviving the negative hypergeometric model. *The American Statistician*, 61(4), 347-350.

Murthy, M.N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhyā*, 18, 379-390.

Ohlsson, E. (1995). Sequential Poisson sampling. Research report 182, Stockholm University, Sweden.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14, 149-162.

Pathak, P.K. (1964). On inverse sampling with unequal probabilities. *Biometrika*, 51, 185-193.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Salehi, M.M., and Seber, G.A.F. (2001). A new proof of Murthy's estimator which applies to sequential sampling. *The Australian and New Zealand Journal of Statistics*, 43, 281-286.

Sampford, M.R. (1962). Methods of cluster sampling with and without replacement for clusters of unequal sizes. *Biometrika*, 49(1/2), 27-40.

Tillé, Y. (1996). An elimination procedure of unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 297-305
Statistics Canada, Catalogue No. 12-001-X

A cautionary note on Clark Winsorization

Mary H. Mulry, Broderick E. Oliver, Stephen J. Kaputa and Katherine J. Thompson[1]

Abstract

Winsorization procedures replace extreme values with less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization therefore both detects and treats influential values. Mulry, Oliver and Kaputa (2014) compare the performance of the one-sided Winsorization method developed by Clark (1995) and described by Chambers, Kokic, Smith and Cruddas (2000) to the performance of *M*-estimation (Beaumont and Alavi 2004) in highly skewed business population data. One aspect of particular interest for methods that detect and treat influential values is the range of values designated as influential, called the detection region. The Clark Winsorization algorithm is easy to implement and can be extremely effective. However, the resultant detection region is highly dependent on the number of influential values in the sample, especially when the survey totals are expected to vary greatly by collection period. In this note, we examine the effect of the number and magnitude of influential values on the detection regions from Clark Winsorization using data simulated to realistically reflect the properties of the population for the Monthly Retail Trade Survey (MRTS) conducted by the U.S. Census Bureau. Estimates from the MRTS and other economic surveys are used in economic indicators, such as the Gross Domestic Product (GDP).

Key Words: Outlier; Masking; Monthly retail trade survey.

1 Introduction

Recently we studied methods of detecting and treating verified influential values with the goal of finding an objective method for identification and treatment of influential values in a highly skewed business population (Mulry et al. 2014). An observation is considered influential if its value is correct but its weighted contribution has an excessive effect on the estimated total or period-to-period change. Although influential values occur infrequently in economic surveys, if one appears and is not "treated," it may introduce substantial over- or under-estimation of survey totals or period-to-period change. In turn, this can impact other measures of the economy. For example, monthly estimates of sales and inventories from the U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) are inputs to the Gross Domestic Product (GDP). With any outlier detection and treatment method, one aspect of particular interest is the range of values that methods designate as influential, called the detection region. The size of the region and its boundary directly impact the number of identified values and the minimum amount by which the value(s) will be adjusted. Consequently, it is important to understand how to "manipulate" the method used, to ensure that (1) true influential values are always identified and receive the minimum treatment needed to ameliorate their impact on totals without overly perturbing the sample's distribution and (2) values that are not influential are rarely identified and are consistently associated with trivial adjustments.

One approach for detecting and treating influential values is called Winsorization. These procedures replace extreme values with other, less extreme values, effectively moving the original extreme values toward the center of the distribution. Winsorization procedures may be one-sided by treating only extreme values that are too high, or they may be two-sided by simultaneously treating high and low values. Values

1. Mary H. Mulry, Broderick E. Oliver, Stephen J. Kaputa, and Katherine J. Thompson, U.S. Census Bureau, Washington, DC 20233, U.S.A. E-mail: mary.h.mulry@census.gov.

designated as influential are modified ("treated") by replacing them with values chosen to minimize the mean squared error (MSE) of the estimate of the total. For further discussion, see Chambers (1986), Chambers et al. (2000), and Martinoz, Haziza and Beaumont (2015).

In this note, we focus on the Clark Winsorization, a one-sided method developed by Clark (1995) and described by Chambers et al. (2000). The Clark Winsorization method assumes a data model and then uses an algorithm to detect and treat influential values. The detected and treated values form the detection region. Our studies found the Clark Winsorization algorithm can be effective, but the resultant detection region is highly dependent on the number of influential values in the sample. If the sample contains no influential values, the procedure is anti-conservative, meaning it makes very small changes to several values not considered influential thus reducing the variance and mean square error but essentially leaving the estimated total unchanged (trimming). On the other hand, the procedure can become very conservative if the sample contains a single influential value, depending on the distance of the value from the remainder of the distribution. When the sample contains two or more influential values, Clark Winsorization detects and adjusts only the influential values and does not trim any values that are not influential. However, the procedure can be prone to masking (Barnett and Lewis 1994). Trimming observations when no influential value is present does not appeal to subject matter analysts in a production setting where time is limited. The cost of examining a "false positive" can be prohibitive and treated values might be categorized as imputed in response rate computations. However, the algorithm has the advantage of being straightforward to implement and not requiring prior knowledge of the population. Certainly there are situations where these advantages of Clark Winsorization may outweigh the disadvantages.

We examine the influential value detection regions from Clark Winsorization using a simulated dataset that realistically reflects the population of the MRTS and was first used in (Mulry et al. 2014). We illustrate how the presence of one versus two high influential values can affect the detection region under several scenarios. Our objective is *not* to advocate for or against this method; the purpose of this note is to make potential users aware of aspects of this procedure that can affect its outcome.

Section 2 contains background on monthly business surveys including an overview of the sample design and weighting. A description of the Clark Winsorization methodology and its implementation using MRTS data appears in Section 3. The discussion in Section 4 concentrates on the detection region for influential values with Section 4.1 addressing the scenario of one influential value in a sample and Section 4.2 focusing on the scenario when two influential values are present. Section 5 contains a summary.

2 Business survey setting

As is typical of many business surveys, the MRTS is sampled from a highly skewed population of companies. The MRTS selects a sample every five years using a stratified simple-random sample design. Primary strata are determined by major industry as reported by the company, whose units are further sub-stratified by estimated annual sales (U.S. Census Bureau 2014). When the sample is introduced, small businesses are generally sampled at a low rate and have large sampling weights, whereas the larger businesses are sampled at a higher rate and have small sampling weights, again typical of many business

survey designs (Smith 2013). Originally, all businesses in the same sampling strata have the same sampling weight. However, weights for individual businesses may be adjusted as the sample matures due to persistent increases in sales for some businesses and decreases for others. For this reason, simulating a realistic weighting structure for a matured sample is challenging. When influential values do appear, it is the combination of the weight and the reported sales that produces the unusually large weighted value.

Sampling weights for small units can be very large, so examining the unweighted values to identify influential values would be quite misleading. We illustrate the combined effect of weight and sales with a single sample (replicate) throughout this note. Figure 2.1 presents plots of sampling weights against unweighted and weighted values of sales, respectively from this sample. Certainty businesses – those sampled with probability equal to 1 – are marked by hollow circles. The graph on the left shows that the units that have the smallest values of sales tend to have the highest sampling weights. By design, as the observed (unweighted) value of sales increases, the sampling weight likewise decreases. However, as shown in the graph on the right, the weighted value from both the small and large businesses contribute similarly to the estimated total. Indeed, a small value of sales multiplied by a large sampling weight can easily affect the estimated total.

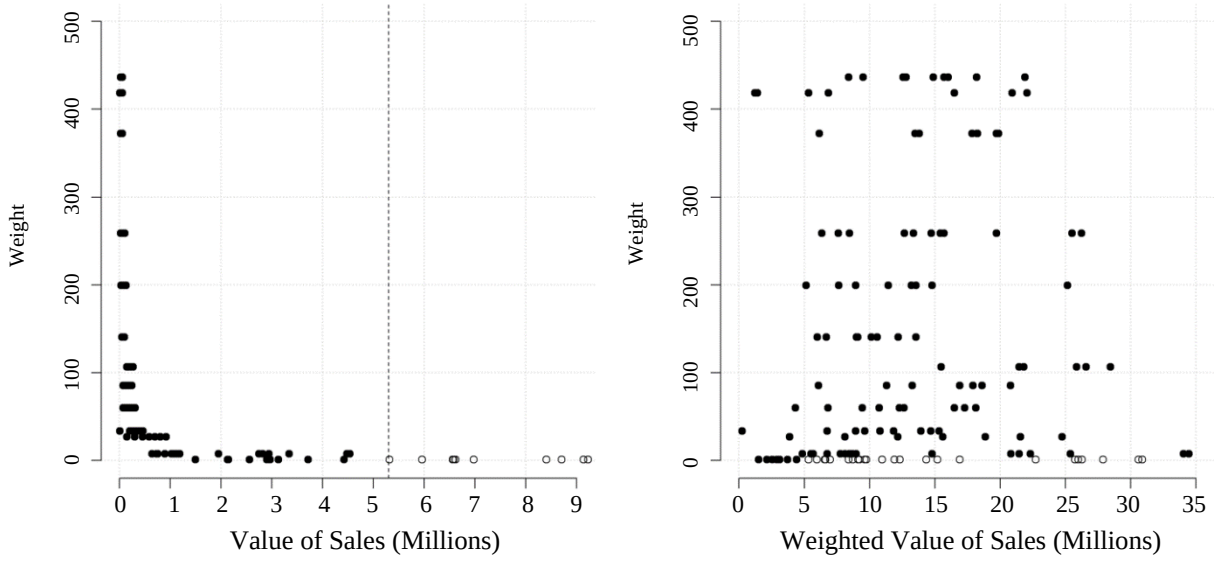

Figure 2.1 On the left, sampling weight versus *unweighted* value of sales. On the right, sampling weight versus *weighted* value of sales for unit. Units selected with certainty are shown as hollow circles.

Economic surveys publish totals and period-to-period change estimates. Influential values are examined with respect to their weighted impact on the total. If the estimates of total sales vary greatly by period, the change estimates are affected accordingly. Currently, when an influential value is detected, the mitigation strategy depends on whether the subject matter experts believe the observation is a one-time phenomenon or a persistent shift. If the influential value appears to be an atypical occurrence for the business, then the

influential observation is replaced with an "imputed" value that is more consistent with the remainder of the distribution whether or not it fails an edit [Note: if the replacement value were obtained via Clark-Winsorization, then it would technically be an "adjusted" value]. If the influential value persists, indicating a permanent change, then its sampling weight is adjusted.

3 Method

We first introduce notation needed to describe the Clark Winsorization, which follows Mulry et al. (2014). For the i^{th} business in a survey sample of size n for the month of observation t, Y_{ti} is the collected characteristic (e.g., sales), w_{ti} is its survey weight (which may or may not be equivalent to the inverse probability of selection), and X_{ti} is a variable highly correlated with Y_{ti}, such as previous month's revenue. The monthly total Y_t is estimated by $\hat{Y}_t$ defined by $\hat{Y}_t = \sum_{i=1}^{n} w_{ti} Y_{ti}$.

For ease of notation, we suppress the index for the month of observation t in the remainder of this section. In MRTS, the survey weight w_i is the (possibly modified) sampling weight since the missing data treatment is imputation.

The general form of the one-sided Winsorized estimator of the total is designed for large values and is written as $\hat{Y}^* = \sum_{i=1}^{n} w_i Z_i$ where $Z_i = \min\{Y_i, K_i + (Y_i - K_i)/w_i\}$.

Detection of observation i as an influential value by Clark Winsorization occurs when $Z_i \neq Y_i$. More than one observation may be identified. Note that using $Z_i = \min\{Y_i, K_i\}$ would ensure bounded influence and a robust estimator. However, this may lead to a large bias in $\hat{Y}^*$.

To implement the method, Clark assumes a general model where the Y_i are characterized as independent realizations of random variables with $E(Y_i) = \mu_i$ and $\text{var}(Y_i) = \sigma_i^2$. Then the approach approximates the K_i that minimizes the MSE under the model by setting $K_i = \mu_i + L(w_i - 1)^{-1}$, which requires estimating μ_i and L. Clark's approach builds on a method developed by Kokic and Bell (1994) that derived a K for each stratum rather than for each individual unit.

For an estimate of μ_i, Chambers et al. (2000) suggest using the results of a robust regression. In our application, we used the SAS Procedure ROBUSTREG (SAS 2014) to implement the weighted least median of squares (LMS) robust regression method. The LMS robust regression uses weights to compensate for the heteroscedasticity visible in Figure 2.1. Other considered methods appeared too sensitive with our data, designating some observations as influential when they were not large enough to have an excessive effect on the estimated total in our empirical data sets. In different applications, different robust regression methods could exhibit superior performance and should be considered. Our prediction model estimates μ_i with bX_i where b is the regression coefficient and X_i is the previous month's observation, chosen because X_i and Y_i tend to be highly correlated and no administrative data are available on a monthly basis. To estimate L, the Clark Winsorization procedure uses the estimate of μ_i to estimate weighted residuals

$$D_i = (Y_i - \mu_i)(w_i - 1) \text{ by } \hat{D}_i = (Y_i - bX_i)(w_i - 1).$$

Certainty units have weighted residual values of zero, assuming that no other weight adjustments are performed (e.g., for unit nonresponse, for post-stratification). Next, the method sorts the estimates of the residuals in *decreasing* order $\hat{D}_{(1)}, \hat{D}_{(2)}, \ldots, \hat{D}_{(n)}$. Then the Clark method finds the largest value of k, called k^*, such that $(k+1)\hat{D}_{(k)} - \sum_{j=1}^{k} \hat{D}_{(j)}$ is positive, then estimates L by $\hat{L} = (k^*+1)^{-1} \sum_{j=1}^{k^*} \hat{D}_{(j)}$. Finally, the estimate of K_i is formed by $\hat{K}_i = bX_i + \hat{L}(w_i - 1)^{-1}$, which is used to determine the values of Z_i for the estimate of the total $\hat{Y}^*$. Chambers et al. (2000) recommend forming the estimate of L for the procedure by using an average of estimates of L from several previous months of data. However, our examples in Section 4 use only the previous month because we use data from a simulated *stationary series* constructed to reflect the different means and variances in the sampling strata for an industry in the MRTS. The stationary series was created by constructing a simulated population from MRTS data and applying an ARMA model to generate the time series. Thus, additions and deletions to the MRTS sample over time (i.e., births and deaths) are not incorporated in the simulation design. Consequently, averaging over several previous months offers no advantage over the point estimate from the previous month. In addition, we used the Winzorized values as auxiliary values (X_i) in the application of the procedure to the subsequent month in order to study the propagation of the effects of the adjustment in the production setting. Although influential values were induced by adding a large amount to an observation selected at random from a stratum with one of the largest weights, the calculation of the value of L used all the sample observations with weights greater than one. More details on the construction of the series may be found in Mulry et al. (2014). We have not explored using an average of estimates of L from several previous months with simulated MRTS data that incorporated seasonality, volatility, and changes in economic conditions or with empirical MRTS data. Such an average of estimates of L may be useful in other designs and surveys that exhibit more stable behavior, such as annual rather than monthly implementations.

4 Detection regions

We examine the range of influential values that Clark Winsorization designates as influential, called the detection region, under three scenarios. One scenario has a single high influential value present in the sample. In the other two scenarios, the sample contains two high influential values.

Figures 4.1 and 4.2 use grids of *unweighted* data to illustrate the detection regions for the application of the Clark Winsorization algorithm on a single sample from a simulated MRTS industry with low volatility, a monthly revenue of $2.5 billion and a sample size of 147. In these figures, each (x, y) point on the grid represents a possible influential value where x represents the unweighted value for the previous month and y represents the current month's unweighted value. Since the weights for the same business rarely change from month-to-month, the scatterplots of weighted values are similar and therefore are not shown. We use sampling weights for the points on the grid and do not modify the weights in our simulation. All the points on a *vertical* line have the same weight with the sample weights lower for units that have higher values of sales. The detection regions are constructed by inserting each pair of (x, y) coordinates from the grid into the sample and then running the Clark Winsorization algorithm with the parameter settings described in Section 3 to see if the weighted y value in the inserted pair is designated as influential.

4.1 Results for one influential value

In this section, we illustrate the effect on the detection region of a sample containing a single influential value, hereafter referred to as *Scenario 1*. In Figure 4.1, the unweighted sample observations used to form the detection regions are shown in black with the x-axis representing the previous month's value and the y-axis representing the current month. The robust least median of squares regression line used in the prediction model has been included for reference. For the given sample, a single observation that falls in the light gray hashed region (detection region) is flagged as influential and adjusted by the Clark Winsorization method. The broken vertical line marks the largest sampled observation with a weight greater than one; that is, all observations to the right of this asymptote are guaranteed to have a weight of one.

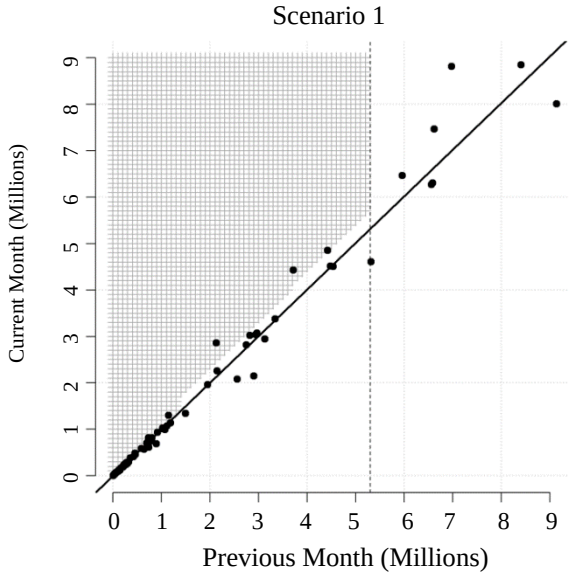

Figure 4.1 Detection region for Clark Winsorization for a single influential value. All sample points are in black.

The close proximity of the lower boundary line of the detection region to the regression line reflects the trimming done by the method to minimize the MSE by lowering the variance at the cost of introducing a small bias. Therefore, several non-influential cases in this detection region will be trimmed slightly nonetheless. We observed this phenomenon repeatedly in several other (different) empirical data sets.

4.2 Results for two influential values

Now we turn to investigating the detection region when the sample contains two induced high influential values. Our approach holds one induced observation at a fixed value and weight in the sample and allows the second induced observation to vary in value with its corresponding weight, permitting the identification of the detection region for the second observation, conditional on the first. This approach allows us to assess whether the procedure is subject to *masking* which occurs when a large value prevents the identification of other extreme values. We consider two scenarios for the fixed value. In *Scenario 2*, the contribution of the

fixed influential value to the estimate of total sales is 667 million higher than the previous month. In *Scenario 3*, the fixed influential value is less severe since its contribution is 334 million higher, half of the increase in *Scenario 2*.

The graph on the left in Figure 4.2 presents the detection region (light gray area) under *Scenario 2*. Here, the fixed (unweighted) value is 350,000 in the previous month and 8.2 million in the current month with a weight of 85. Regardless of where the second observation was placed throughout the graph, the fixed observation was always designated as influential. Notice that the observations that would have been *falsely* designated as influential and slightly trimmed in *Scenario 1* (see Figure 4.1) would *not* have been changed in this scenario. Here, the detection region is restricted to identifying only *similar* severe observations, which are supposed to be atypical.

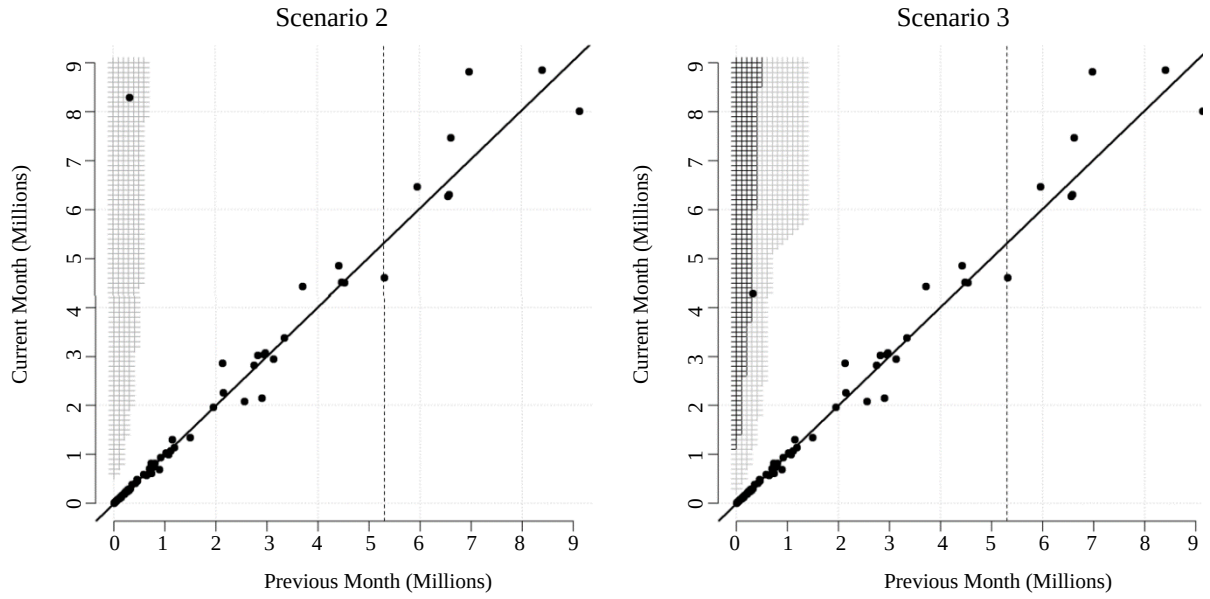

Figure 4.2 Detection regions for Clark Winsorization for the second of two influential values. On the left, the first one is held fixed and is extreme. On the right, the first one is held fixed but is less extreme. All sample points are in black.

The dramatic difference in the relative sizes of the detection regions between *Scenario 1* and *Scenario 2* could indicate that this procedure – as applied – is vulnerable to masking. Masking occurs when one influential value causes a failure to identify the presence of another (Barnett and Lewis 1994). We explore this possibility in *Scenario 3*, halving the unweighted value of the fixed influential value in the current month (now 4.1 million instead of 8.2 million) while allowing the weight to retain the same value of 85. The graph on the right in Figure 4.2 shows two different shaded regions: a light grey area where *both* the fixed influential value and the second (variable) value can be detected, and a dark gray region to the left of the light gray region where the algorithm detects the *variable value* as influential but misses the fixed value.

Adjustments in the light gray area reduce both the bias and the MSE. In the dark gray area, the adjustment reduces the MSE, but may not reduce the bias substantially. The white area to the right of the light gray region shows where only the fixed influential value is identified. However, the white area contains large observations with small weights so these observations are not representing much more than themselves, and consequently adjustments in this range have small impact on the bias.

This preliminary exploration validates our concern about the potential for masking. One approach that may alleviate masking when a stationary series has a high level of noise is to average L over several previous months as suggested in Chambers et al. (2000). The sampling design may be a factor. The graph on the left in Figure 2.1 shows that the weights decline rapidly as the unweighted observations increase for observations between 0 and 1 million. In this range, the weight of the unit has more impact than its observed value on the size of its weighted residual used in calculating the k^*. A relatively small change in the variable value may trigger a much larger change in its weighted residual and cause the k^* to change, which affects the number of influential values detected. The weights used in this example reflect the weights used in the MRTS for the industry and were not constructed artificially to create an illustration for the Clark Winsorization methodology.

5 Summary

The usage of Clark Winsorization is very appealing for the simplicity of its implementation and lack of parameters as long as one can build a viable robust regression model. However, as with many outlier detection procedures, the method has certain vulnerabilities that are not always obvious. This note demonstrates how the procedure can be effective at identifying and treating influential values, but is also highly sensitive to the number of influential values in the sample and their magnitude with respect to the regression line used to determine the detection region bounds. The properties of the detection region vary by whether an influential value is present and by the number and severity when one or more appear. If the sample contains no influential values, the procedure is anti-conservative in that it trims values not considered influential to minimize the MSE (by reducing the variance). In contrast, the procedure can become very conservative depending on the degree of difference of the weighted influential value from the others in the sample. When the sample contains two or more influential values, Clark Winsorization detects and adjusts only the influential values and does not trim any values that are not influential. However, our results demonstrate a potential for masking which should be considered when implementing the procedure.

If the occurrence of an influential value is truly a rare event and large influential values are of interest, then the small trimming of a handful of values that are not influential is a disadvantage. However, in applications where influential values are common or where historic data are not available for modeling, implementing Clark Winsorization definitely requires an assessment of the amount of trimming to determine if the aggregated small changes greatly affect the estimated total. If not, then this is an appealing approach. If yes, then other methods such as *M*-estimation – which give more control over the detection region – may be advantageous.

Acknowledgements

This report is released to inform interested parties and encourage discussion of work in progress. The views expressed on statistical, methodological, and operational issues are those of the authors and not necessarily those of the U.S. Census Bureau. The authors thank Lynn Weidman, Eric Slud, Scott Scheleur, William C. Davie Jr. and Carma Hogue for their helpful reviews of previous versions of the manuscript. The authors also thank Ray Chambers for his comments during presentations of our work in progress. The authors appreciate the comments from the Associate Editor and the anonymous Referees.

References

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data. 3rd Edition.* New York: John Wiley & Sons, Inc.

Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 2, 195-208. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2004002/article/7752-eng.pdf.

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chambers, R., Kokic, P., Smith, P. and Cruddas, M. (2000). Winsorization for identifying and treating outliers in economic surveys. *ICES II, The Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions*, American Statistical Association, 717-726.

Clark, R. (1995). *Winsorization Methods in Sample Surveys*. Masters Thesis. Department of Statistics. Australia National University. http://hdl.handle.net/10440/1031 (accessed September 29, 2016.).

Kokic, P.N., and Bell, P.A. (1994). Optimal winsorising cut-offs for a stratified finite population estimator. *Journal of Official Statistics*, Stockholm, Sweden, 10, 419-435.

Martinoz, C.F., Haziza, D. and Beaumont, J.-F. (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Survey Methodology*, 41, 1, 57-77. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015001/article/14199-eng.pdf.

Mulry, M.H., Oliver, B.E. and Kaputa, S.J. (2014). Detecting and treating verified influential values in a monthly retail trade survey. *Journal of Official Statistics*, 30(4), 1-28.

SAS (2014). *Help and Documentation*. SAS Institute, Inc. Cary, NC.

Smith, P. (2013). Sampling and estimation for business surveys. In *Designing and Conducting Business Surveys*, (Eds., G. Snijkers, G. Haraldsen, J. Jones and D. Willimack), Hoboken, NJ: John Wiley & Sons, Inc., 219-52.

U.S. Census Bureau (2014). *Monthly Retail Trade Survey Methodology*. U.S. Census Bureau, Washington, DC. http://www.census.gov/retail/mrts/how_surveys_are_collected.html (accessed September 29, 2016).

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 307-317
Statistics Canada, Catalogue No. 12-001-X

A few remarks on a small example by Jean-Claude Deville regarding non-ignorable non-response

Yves Tillé[1]

Abstract

An example presented by Jean-Claude Deville in 2005 is subjected to three estimation methods: the method of moments, the maximum likelihood method, and generalized calibration. The three methods yield exactly the same results for the two non-response models. A discussion follows on how to choose the most appropriate model.

Key Words: Calibration; Generalized calibration; Method of moments; Likelihood.

1 Deville's example

During a conference at the University of Neuchâtel, Jean-Claude Deville (2005) presented a simple example to illustrate the value of generalized calibration for dealing with non-ignorable non-response (regarding generalized calibration, see Deville 2000, 2002 and 2004; Kott 2006; Chang and Kott 2008; Kott and Chang 2010; and Lesage and Haziza 2015). The example is reproduced below in its entirety.

> *Adjustments to offset the effects of non-response require very accurate knowledge of the factors that cause it. In particular, if what is to be measured directly influences the response probability, we must take risks with the data. Here is a small fictional example: A group of students is interviewed about their use of drugs. The survey results are as follows:*

Table 1.1
Deville's example

	YES	NO	NON-RESPONSE	COMBINED
Boys	40	80	180	300
Girls	20	160	120	300
Combined	60	240	300	600

> *Naively, we would think that the percentage of drug users is estimated at 60/(240 + 60) = 25%. This estimate is made under the assumption that non-respondents have the same behaviour as respondents. However, we notice that the response rate for girls is greater than the response rate for boys. To correct that, we calculate the rate of drug users among girls, or 1/9, and among boys, or 3/9, and we conclude that the rate of drug users in observed student population is 2/9 = 22.2%. Now, if we think that drug use is causing the non-response, the model has two*

1. Yves Tillé, Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland. E-mail: yves.tille@unine.ch.

parameters p_{yes} and p_{no}, the response probabilities of users and non-users, respectively. We find that these probabilities equal 0.2 and 0.8, respectively. The estimated number of users is therefore 200 among boys and 100 among girls, and the estimated overall percentage is 50!

At first glance, the example is simple, and it perfectly explains the usual typology of the three non-response mechanisms. Each of the three estimates proposed in the example corresponds to one of the three categories below:

- Missing completely at random (MCAR): The response probability does not depend on the variable of interest (drug use) or on the auxiliary variable (gender).
- Missing at random (MAR): The response probability does not depend on the variable of interest y after conditioning on the auxiliary variable x (gender). In this case, the response probability would therefore depend on gender only.
- Not missing at random (NMAR): The response probability depends on the variable of interest itself (drug use) even if consideration is given to the auxiliary variable x.

The example shows the value of generalized calibration, which can deal directly with NMAR. Jean-Claude Deville addresses the problem by considering the probabilities p_{yes} and p_{no} as parameters to be estimated. This example can be dealt with in several ways, depending on one's point of view on inference.

In the following, we will show that there are at least three methods to address the problem, namely the method of moments, the maximum likelihood method and calibration. The maximum likelihood method was not dealt with by Jean-Claude Deville. We develop calculations completely for the first two estimation methods by considering the two models. We also calculate the calibration and generalized calibration results.

We show that the three results obtained are identical. The estimated likelihood function could be used to choose between the two models. Unfortunately, the function has the same value for both models, which does not make it possible to choose the model. However, we propose a way to make a choice.

In Section 2, we present the notation used. Section 3 is devoted to estimation using the method of moments, and Section 4 is devoted to estimation using the maximum likelihood method. In Section 5, we apply the calibration and generalized calibration methods. We close with a discussion on the value of each method in Section 6.

2 Notation

Table 2.1 shows the notation for Table 1.1.

Table 2.1
Notation for Table 1.1

	Drug User	Non-user	Missing	Total
Male	r_{HD}	r_{HS}	m_H	$n_{H.}$
Female	r_{FD}	r_{FS}	m_F	$n_{F.}$
Total	$r_{.D}$	$r_{.S}$	m	n

For simplicity, assume that we are dealing with a census. In other words, the 600 students were not randomly selected. Therefore, the only source of randomness is the non-response mechanism. This assumption is not that restrictive, since it is equivalent to considering that the sample is random, but that the reasoning below is conditional on the random sample. The objective is to estimate the numbers of people in Table 2.2. This table is assumed not to be random. It is therefore a matter of distributing the non-respondents m_H and m_F between drug users and non-users.

Table 2.2
Number of people to be estimated based on Table 1.1

	Drug User	Non-user	Total
Male	n_{HD}	n_{HS}	$n_{H.}$
Female	n_{FD}	n_{FS}	$n_{F.}$
Total	$n_{.D}$	$n_{.S}$	n

As well, it is assumed that the non-response follows a Poisson design, that is, each individual decides whether or not to respond with a probability independent of other individuals. The response probability may vary among individuals.

The two vectors (r_{HD}, r_{HS}, m_H), and (r_{FD}, r_{FS}, m_F) each have a multinomial distribution whose parameters depend on the model used. MCAR cases, which are completely trivial, will not be studied. In Table 2.3, which shows cases of MAR, the response probability depends on gender only (p_H for males, p_F for females). In Table 2.4, which shows cases of NMAR, the response probability depends only on being or not being a drug user (q_D, q_S for the others).

Table 2.3
Case 1: MAR model, non-response depends on gender

	Drug User	Non-user	Missing	Total
Male	$\mathrm{E}(r_{HD}) = n_{HD}p_H$	$\mathrm{E}(r_{HS}) = n_{HS}p_H$	$\mathrm{E}(m_H) = n_{H.}(1-p_H)$	$n_{H.}$
Female	$\mathrm{E}(r_{FD}) = n_{FD}p_F$	$\mathrm{E}(r_{FS}) = n_{FS}p_F$	$\mathrm{E}(m_F) = n_{F.}(1-p_F)$	$n_{F.}$
Total	$\mathrm{E}(r_{.D})$	$\mathrm{E}(r_{.S})$	m	n

Table 2.4
Case 2: NMAR model, non-response depends on being or not being a drug user

	Drug User	Non-user	Missing	Total
Male	$\mathrm{E}(r_{HD}) = n_{HD}q_D$	$\mathrm{E}(r_{HS}) = n_{HS}q_S$	$\mathrm{E}(m_H) = n_{HD}(1-q_D) + n_{HS}(1-q_S)$	$n_{H.}$
Female	$\mathrm{E}(r_{FD}) = n_{FD}q_D$	$\mathrm{E}(r_{FS}) = n_{FS}q_S$	$\mathrm{E}(m_F) = n_{FD}(1-q_D) + n_{FS}(1-q_S)$	$n_{F.}$
Total	$\mathrm{E}(r_{.D})$	$\mathrm{E}(r_{.S})$	m	n

3 Estimation using the method of moments

3.1 MAR

The method of moments makes it possible to estimate parameters quickly. For MAR, we obtain the third column of Table 2.3 using the equations

$$\begin{aligned} \mathrm{E}(m_H) &= n_{H.}(1-p_H), \\ \mathrm{E}(m_F) &= n_{F.}(1-p_F), \end{aligned}$$

which yield the estimators

$$\begin{aligned} \hat{p}_H &= 1-\frac{m_H}{n_{H.}}, \\ \hat{p}_F &= 1-\frac{m_F}{n_{F.}}, \end{aligned}$$

and therefore, from the first two columns,

$$\begin{aligned} \hat{n}_{.D} &= \frac{r_{HD}}{\hat{p}_H}+\frac{r_{FD}}{\hat{p}_F} = r_{HD}\frac{n_{H.}}{n_{H.}-m_H}+r_{FD}\frac{n_{F.}}{n_{F.}-m_F}, \\ \hat{n}_{.S} &= \frac{r_{HS}}{\hat{p}_H}+\frac{r_{FS}}{\hat{p}_F} = r_{HS}\frac{n_{H.}}{n_{H.}-m_H}+r_{FS}\frac{n_{F.}}{n_{F.}-m_F}. \end{aligned}$$

The estimated response probabilities are $\hat{p}_H = 0.4$ and $\hat{p}_F = 0.6$. We therefore obtain the estimates shown in Table 3.1.

Table 3.1
Estimates: MAR

	YES	NO	COMBINED
Boys	100.00	200.00	300
Girls	33.33	266.66	300
COMBINED	133.33	466.66	600

3.2 NMAR

For NMAR, we obtain the following equations from Table 2.4:

$$\begin{aligned} \mathrm{E}(m_H) &= E(r_{HD})\frac{1-q_D}{q_D}+\mathrm{E}(r_{HS})\frac{1-q_S}{q_S}, \\ \mathrm{E}(m_F) &= E(r_{FD})\frac{1-q_D}{q_D}+\mathrm{E}(r_{FS})\frac{1-q_S}{q_S}. \end{aligned}$$

After a few calculations, we obtain the following response probability estimators:

$$\hat{q}_D = \frac{r_{HD}r_{FS} - r_{FD}r_{HS}}{(m_H + r_{HD})r_{FS} - (m_F + r_{FD})r_{HS}},$$

$$\hat{q}_S = \frac{r_{HD}r_{FS} - r_{FD}r_{HS}}{(m_F + r_{FS})r_{HD} - (m_H + r_{HS})r_{FD}}.$$

Finally, we obtain

$$\hat{n}_{.D} = \frac{r_{.D}}{\hat{q}_D} = r_{.D}\frac{(m_H + r_{HD})r_{FS} - (m_F + r_{FD})r_{HS}}{r_{HD}r_{FS} - r_{FD}r_{HS}} = r_{.D}\frac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{HD}r_{FS} - r_{FD}r_{HS}},$$

$$\hat{n}_{.S} = \frac{r_{.S}}{\hat{q}_S} = r_{.S}\frac{(m_F + r_{FS})r_{HD} - (m_H + r_{HS})r_{FD}}{r_{HD}r_{FS} - r_{FD}r_{HS}} = r_{.S}\frac{n_{F.}r_{HD} - n_{H.}r_{FD}}{r_{HD}r_{FS} - r_{FD}r_{HS}}.$$

As Deville writes, the estimated response probabilities are $\hat{q}_D = 0.2$ and $\hat{q}_S = 0.8$. We therefore obtain the estimates in Table 3.2.

Table 3.2
Estimates: NMAR

	YES	NO	COMBINED
Boys	200	100	300
Girls	100	200	300
COMBINED	300	300	600

4 Estimation using the maximum likelihood method

4.1 MAR

The probability distribution is multinomial. For MAR, the following likelihood function applies:

$$\mathcal{L}(n_{HD}, n_{FD}, p_H, p_F) = \frac{n_{H.}!}{r_{HD}!\, r_{HS}!\, m_H!}\left(\frac{n_{HD}p_H}{n_{H.}}\right)^{r_{HD}}\left(\frac{(n_{H.} - n_{HD})p_H}{n_{H.}}\right)^{r_{HS}}\left(\frac{n_{H.}(1-p_H)}{n_{H.}}\right)^{m_H}$$
$$\times\frac{n_{F.}!}{r_{FD}!\, r_{FS}!\, m_F!}\left(\frac{n_{FD}p_F}{n_{F.}}\right)^{r_{FD}}\left(\frac{(n_{F.} - n_{FD})p_F}{n_{F.}}\right)^{r_{FS}}\left(\frac{n_{F.}(1-p_F)}{n_{F.}}\right)^{m_F}.$$

By setting to zero the partial derivatives of the log-likelihood with respect to parameters p_H and p_F, we obtain two equations with two unknowns. The solution yields the estimators

$$\hat{p}_H = 1 - \frac{m_H}{n_{H.}},$$

$$\hat{p}_F = 1 - \frac{m_F}{n_{F.}}.$$

By setting to zero the derivatives with respect to n_{HD} and n_{FD}, we obtain the estimators

$$\hat{n}_{HD} = \frac{r_{HD}}{\hat{p}_H} \text{ and } \hat{n}_{FD} = \frac{r_{FD}}{\hat{p}_F}.$$

Therefore,

$$\hat{n}_{.D} = \hat{n}_{HD} + \hat{n}_{FD} = \frac{r_{HD}}{\hat{p}_H} + \frac{r_{FD}}{\hat{p}_F}.$$

These estimators are exactly the same as those obtained using the method of moments.

4.2 NMAR

For NMAR, the following likelihood function applies:

$$\mathcal{L}(n_{HD}, n_{FD}, q_D, p_S) = \frac{n_{H.}!}{r_{HD}!\, r_{HS}!\, m_H!}\left(\frac{n_{HD} q_D}{n_{H.}}\right)^{r_{HD}}\left(\frac{(n_{H.} - n_{HD})\, q_S}{n_{H.}}\right)^{r_{HS}}\left(\frac{n_{HD}(1-q_D) + (n_{H.} - n_{HD})(1-q_S)}{n_{H.}}\right)^{m_H}$$
$$\times \frac{n_{F.}!}{r_{FD}!\, r_{FS}!\, m_F!}\left(\frac{n_{FD} q_D}{n_{F.}}\right)^{r_{FD}}\left(\frac{(n_{F.} - n_{FD})\, q_S}{n_{F.}}\right)^{r_{FS}}\left(\frac{n_{FD}(1-q_D) + (n_{F.} - n_{FD})(1-q_S)}{n_{F.}}\right)^{m_F}.$$

By setting to zero the partial derivatives of the log-likelihood with respect to the four parameters q_D, q_S, n_{HD} and n_{FD}, we obtain a system of four rather complicated second-order equations with four unknowns. We used a symbolic computation software program to verify that the solution given by the method of moments is a solution to this system of equations. Obviously, since the system is second-order, there is a second solution. However, for Deville's example, the second solution yields negative values, which are not valid for estimating probabilities and numbers of people.

5 Estimation using calibration and generalized calibration

5.1 Notation

To define calibration, we will establish the following notation. Let $U = \{1, \ldots, k, \ldots, N\}$ be the set of people interviewed (here, $N = 600$) and $R \subset U$ be the set of respondents to the question regarding drug use. As well, we define the following:

$$\mathbf{x}_k = \begin{cases} (1 \quad 0)^{\mathrm{T}} & \text{if individual } k \text{ is male} \\ (0 \quad 1)^{\mathrm{T}} & \text{if individual } k \text{ is female.} \end{cases}$$

and

$$\mathbf{z}_k = \begin{cases} (1 \quad 0)^{\mathrm{T}} & \text{if individual } k \text{ reported using drugs} \\ (0 \quad 1)^{\mathrm{T}} & \text{if individual } k \text{ reported not using drugs.} \end{cases}$$

Using the notation defined above,

$$\sum_{k\in U}\mathbf{x}_k = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix}, \quad \sum_{k\in R}\mathbf{x}_k = \begin{pmatrix} n_{H.}-m_H \\ n_{F.}-m_F \end{pmatrix}, \quad \sum_{k\in R}\mathbf{z}_k = \begin{pmatrix} r_{.D} \\ r_{.S} \end{pmatrix},$$

$$\sum_{k\in R}\mathbf{x}_k\mathbf{x}_k^{\mathrm{T}} = \begin{pmatrix} n_{H.}-m_H & 0 \\ 0 & n_{F.}-m_F \end{pmatrix}, \quad \sum_{k\in R}\mathbf{x}_k\mathbf{z}_k^{\mathrm{T}} = \begin{pmatrix} r_{HD} & r_{HS} \\ r_{FD} & r_{FS} \end{pmatrix},$$

and

$$\sum_{k\in R}\mathbf{z}_k\mathbf{z}_k^{\mathrm{T}} = \begin{pmatrix} r_{.D} & 0 \\ 0 & r_{.S} \end{pmatrix}.$$

5.2 Estimation using simple calibration

Using simple calibration as described in Deville and Särndal (1992), we seek a weight that is expressed as

$$w_k = F\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right),$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ is a parameter vector and $F(.)$ is a calibration function, that is, a strictly increasing function such that $F(0) = 1$ and whose derivative $F'(.)$ is such that $F'(0) = 1$.

Vector $\boldsymbol{\lambda}$ is determined by using the Newton method to solve the system of equations

$$\sum_{k\in R}F\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k. \tag{5.1}$$

Finally, the calibration estimator is given by

$$\begin{pmatrix} \hat{n}_{.D} \\ \hat{n}_{.S} \end{pmatrix} = \sum_{k\in R} w_k\mathbf{z}_k.$$

In our application, equation (5.1) becomes

$$\sum_{k\in R}F\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \begin{pmatrix} (n_{H.}-m_H)F(\lambda_1) \\ (n_{F.}-m_F)F(\lambda_2) \end{pmatrix} = \sum_{k\in U}\mathbf{x}_k = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix}.$$

We directly obtain the following:

$$w_k = F\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = \begin{cases} n_{H.}/(n_{H.}-m_H) & \text{if individual } k \text{ is male} \\ n_{F.}/(n_{F.}-m_F) & \text{if individual } k \text{ is female.} \end{cases}$$

Therefore, the calibrated estimators are

$$\hat{n}_{.D} = r_{HD}\frac{n_{H.}}{n_{H.}-m_H} + r_{FD}\frac{n_{F.}}{n_{F.}-m_F}$$

$$\hat{n}_{.S} = r_{HS}\frac{n_{H.}}{n_{H.}-m_H} + r_{FS}\frac{n_{F.}}{n_{F.}-m_F},$$

which is exactly the same result as that yielded by the method of moments and the maximum likelihood method. In this case, the solution does not depend on the calibration function used. Obviously, the example is especially simple. In more complex cases where the category definitions do not overlap, the result depends on the calibration function used.

5.3 Generalized calibration

For generalized calibration as defined in (Deville 2000, 2002, 2004; Kott 2006), the weights are expressed as

$$w_k = F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right).$$

Vector $\boldsymbol{\lambda}$ is determined by solving the system of equations

$$\sum_{k\in R} F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k. \tag{5.2}$$

Finally, the generalized calibration estimator is given by

$$\begin{pmatrix}\hat{n}_{.D}\\ \hat{n}_{.S}\end{pmatrix} = \sum_{k\in R} w_k\mathbf{z}_k.$$

In our application, equation (5.2) becomes

$$\sum_{k\in R} F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \begin{pmatrix} r_{HD}F\left(\lambda_1\right) + r_{HS}F\left(\lambda_2\right)\\ r_{FD}F\left(\lambda_1\right) + r_{FS}F\left(\lambda_2\right)\end{pmatrix} = \sum_{k\in U}\mathbf{x}_k = \begin{pmatrix} n_{H.}\\ n_{F.}\end{pmatrix},$$

Which can be written as a matrix

$$\begin{pmatrix} r_{HD} & r_{HS}\\ r_{FD} & r_{FS}\end{pmatrix}\begin{pmatrix} F\left(\lambda_1\right)\\ F\left(\lambda_2\right)\end{pmatrix} = \begin{pmatrix} n_{H.}\\ n_{F.}\end{pmatrix}.$$

We simply solve the linear system

$$\begin{pmatrix} F\left(\lambda_1\right)\\ F\left(\lambda_2\right)\end{pmatrix} = \begin{pmatrix} r_{HD} & r_{HS}\\ r_{FD} & r_{FS}\end{pmatrix}^{-1}\begin{pmatrix} n_{H.}\\ n_{F.}\end{pmatrix} = \begin{pmatrix} \dfrac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{FS}r_{HD} - r_{FD}r_{HS}}\\ \dfrac{n_{H.}r_{FD} - n_{F.}r_{HD}}{r_{FD}r_{HS} - r_{FS}r_{HD}}\end{pmatrix}.$$

The estimators are therefore

$$\hat{n}_{.D} = r_{.D}\frac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{FS}r_{HD} - r_{FD}r_{HS}}$$

$$\hat{n}_{.S} = r_{.S}\frac{n_{H.}r_{FD} - n_{F.}r_{HD}}{r_{FD}r_{HS} - r_{FS}r_{HD}}.$$

Again, the solution does not depend on the calibration function used. The solution is identical to the solution obtained using the method of moments and the maximum likelihood method. Here, too, this property results from the simplicity of the example. In more complex cases, the result depends on the calibration function used.

6 Discussion

Deville's example is especially welcome since, for both models, the three estimation methods provide exactly the same estimators. Obviously, if the model is more complicated, using the maximum likelihood method becomes cumbersome, if not impossible. The calibration and generalized calibration method works in all cases as long as the number of calibration variables whose totals are known is sufficient and the matrix

$$\sum_{k \in R} \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}}$$

is invertible. In this example, the determinant of this matrix appears in the denominator of the estimators. Therefore, a small determinant makes the estimates especially risky. Lesage and Haziza (2015) recommend verifying that the correlations between variables $\mathbf{x}_k$ and $\mathbf{z}_k$ are great enough to avoid potentially amplifying the bias.

If the variables are quantitative, the solutions will depend on the calibration function used $F(.)$. The use of the calibration function $F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = 1 + \exp\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)$ is recommended, since it has the advantage of providing weights greater than 1. The inverse of the weights can now be interpreted as a response probability estimated using a logistic model.

The main difficulty is obviously choosing between the two proposed models. In Deville's example, it may seem more "logical" to see the non-response depend rather on drug use than on gender. However, we are not well equipped to make a choice between the two models. The values of the two likelihood functions for the estimated parameters are equal. Is it possible to choose the model based on more than a strong conviction? As suggested in Haziza and Lesage (2016), we recommend always calculating both weightings and comparing the weights and estimates obtained with each of them.

One option may be to calculate an indicator of the dispersion of the response probabilities, such as the variance. For example, if the variance is great, it means that the model has made it possible to calculate response probabilities with greater contrast between individuals and that the model has therefore taken better account of the non-response. Validation through a search for contrasting weights is the basis for identifying response homogeneity groups (RHGs) for all segmentation methods, for example with the chi-square automatic interaction detector (CHAID) algorithm developed by Kass (1980). For example, with CHAID, in each step the RHGs are split based on categories that result in response probabilities with the greatest contrast. By using the same principle in choosing the model, we can select the model that provides the weights with the greatest contrast. For example, if the variance is small, it means that the non-response model could not highlight the differences in non-response probabilities between individuals. Incidentally, the variance in response probabilities is the square of the R-indicator defined by Schouten, Cobben and Bethlehem (2009), used here to choose a non-response model.

In both cases, the average response probability equals 0.5. Specifically,

$$\overline{p} = n_{H.}\frac{n_{H.}\hat{p}_H + n_{F.}\hat{p}_F}{n} = \frac{300 \times 0.4 + 300 \times 0.6}{600} = 0.5$$

and

$$\overline{q} = \hat{n}_{.D}\frac{n_{.D}\hat{q}_D + \hat{n}_{.S}\hat{q}_S}{n} = \frac{300\times 0.2 + 300\times 0.8}{600} = 0.5.$$

For the MAR model, the variance is

$$V_{MAR} = \frac{n_{H.}\left(\hat{p}_H - \overline{p}\right)^2 + n_{F.}\left(\hat{p}_F - \overline{p}\right)^2}{n} = \frac{300(0.4-0.5)^2 + 300(0.6-0.5)^2}{600} = 0.01.$$

For the NMAR model, the variance is

$$V_{NMAR} = \frac{\hat{n}_{.D}\left(\hat{q}_D - \overline{q}\right)^2 + \hat{n}_{.S}\left(\hat{q}_S - \overline{q}\right)^2}{n} = \frac{300(0.2-0.5)^2 + 300(0.8-0.5)^2}{600} = 0.09.$$

The greater variance of the NMAR model is an argument in its favour. In fact, the response probabilities show much greater contrast.

Acknowledgements

The author thanks Audrey-Anne Vallée for her meticulous proofreading of an earlier version of this text and an anonymous referee for their especially pertinent comments.

References

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *Compstat - Proceedings in Computational Statistics:* 14th Symposium held in Utrecht, Netherlands, pages 65-76, New York: Springer.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. In the *Actes des Journées de Méthodologie Statistique*, Paris. Insee-Méthodes.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Technical report, internal document, INSEE, Paris.

Deville, J.-C. (2005). Calibration, past, present and future? Presentation at the conference: *Calibration Tools for Survey Statisticians*, Neuchâtel.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Haziza, D., and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. Will appear in the *Journal of Official Statistics*.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 119-127.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2006002/article/9547-eng.pdf.

Kott, P.S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491), 1265-1275.

Lesage, E., and Haziza, D. (2015). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. Under revision for *Journal of Survey Statistics and Methodology*.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2009001/article/10887-eng.pdf.

Survey Methodology, December 2016
Vol. 42, No. 2, pp. 319-323
Statistics Canada, Catalogue No. 12-001-X

A note on the concept of invariance in two-phase sampling designs

Jean-François Beaumont and David Haziza[1]

Abstract

Two-phase sampling designs are often used in surveys when the sampling frame contains little or no auxiliary information. In this note, we shed some light on the concept of invariance, which is often mentioned in the context of two-phase sampling designs. We define two types of invariant two-phase designs: strongly invariant and weakly invariant two-phase designs. Some examples are given. Finally, we describe the implications of strong and weak invariance from an inference point of view.

Key Words: Double expansion estimator; Horvitz–Thompson estimator; Strong invariance; Two-phase sampling; Weak invariance.

1 Introduction

Two-phase sampling designs are often used in surveys when the sampling frame contains little or no auxiliary information. It consists of first selecting a large sample from the population (typically using a rudimentary sampling design) in order to collect data on variables that are inexpensive to obtain and that are related to the characteristics of interest. The idea behind two-phase sampling is to create a pseudo-sampling frame richer in auxiliary information than the original sampling frame. Then, using the variables observed in the first phase, an efficient sampling procedure can be used to select a (typically small) subsample from the first-phase sample in order to collect the characteristics of interest. Two-phase sampling may also be helpful in a context of nonresponse as the set of respondents is often viewed as a second-phase sample.

We adopt the following notation: consider a population U of size N. A vector $\mathbf{I}_1$ is generated according to the sampling design $F(\mathbf{I}_1)$, where $\mathbf{I}_1 = (I_{11},\ldots,I_{1N})^{\mathrm{T}}$ denotes a vector of indicators such that I_{1i} is either equal to 0 or 1. The first-phase sample, denoted by s_1, is the set of population units for which $I_{1i} = 1$ and $n_1 = \sum_{i\in U} I_{1i}$, is the size of s_1. Then, a vector $\mathbf{I}_2$ is generated according to the sampling design $F(\mathbf{I}_2 \mid \mathbf{I}_1)$, where $\mathbf{I}_2 = (I_{21},\ldots,I_{2N})^{\mathrm{T}}$ denotes the vector of indicators such that I_{2i} is either equal to 0 or 1. The second-phase sample, denoted by s_2 is the set of population units for which both $I_{1i} = 1$ and $I_{2i} = 1$ and $n_2 = \sum_{i\in U} I_{1i}I_{2i}$ is the size of s_2. In practice, note that the indicators I_{2i} are not generated for the population units belonging to the set $U - s_1$. However, at least conceptually, nothing precludes defining these indicators for the units outside the first-phase sample.

Let $\pi_{1i} = P(I_{1i} = 1)$ and $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$ be the first-order and second-order selection probabilities at the first-phase. Similarly, let $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 \mid \mathbf{I}_1)$ and $\pi_{2ij}(\mathbf{I}_1) = P(I_{2i} = 1, I_{2j} = 1 \mid \mathbf{I}_1)$ be the first-order and second-order selection probabilities at the second-phase. Note that the (first-order and second-order) selection probabilities at the second-phase may depend on the realized sample s_1.

1. Jean-François Beaumont, International Cooperation and Corporate Statistical Methods Division, Statistics Canada, 100 Tunney's Pasture Driveway, R.H. Coats Building, 25th floor, Ottawa, Canada, K1A 0T6. E-mail: jean-francois.beaumont@canada.ca; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

The paper is organized as follows. In Section 2, we define the concepts of weak and strong invariance and provide some examples. In Section 3, we discuss the implications of weak and strong invariance from an inferential point of view. In particular, we discuss the reverse decomposition of the variance in the case of a strongly invariant two-phase sampling design.

2 The concept of invariance

We distinguish the concept of strong invariance that may also be called distribution invariance from that of weak invariance that may also be called first-two-moment invariance.

Definition 1. *A two-phase sampling design is said to be strongly (or distribution) invariant provided that*

$$F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2) \tag{2.1}$$

A consequence of Definition 1 is that $F(\mathbf{I}_1, \mathbf{I}_2) = F(\mathbf{I}_1)F(\mathbf{I}_2)$ and therefore, with a strongly invariant two-phase sampling design, the vector $\mathbf{I}_2$ can be generated prior to the vector $\mathbf{I}_1$. In practice, the concept of strong invariance is satisfied for only few two-phase sampling designs. A first example is Poisson sampling at the second phase. This covers the case of nonresponse, which is often viewed as a Poisson sampling design at the second phase. An other example is two-stage sampling. Both are described in greater detail below.

Example 1. *At the first phase, a sample s_1 is selected according to an arbitrary sampling design followed by Poisson sampling at the second phase, where the units selection probability $\pi_{2i}(\mathbf{I}_1)$ are set prior to sampling, which means that $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ for $i \in U$. Since Poisson sampling is completely characterized by its first-order selection probabilities, we have $F(\mathbf{I}_2 \mid \mathbf{I}_1) = F(\mathbf{I}_2)$. As a result, this sampling design is strongly invariant. It can be implemented as follows: first, generate the vector $\mathbf{I}_2$ according to the Poisson sampling design $F(\mathbf{I}_2)$ and, independently, generate the vector $\mathbf{I}_1$ according to the design $F(\mathbf{I}_1)$.*

Example 2. *Two-stage cluster sampling can be described as follows: at the first stage, a sample of clusters is selected randomly from the population of clusters. Then, at the second stage, within each cluster selected at the first stage, a sample of elements is randomly selected. Note that, even in this case, the vector $\mathbf{I}_1$ is still defined at the element level, with its size N corresponding to the number of elements in the population. Under this set-up, the selection indicator for an element j within cluster i, I_{1ij}, is equal to 1 for all elements j within a selected cluster i. Therefore, two-stage sampling is a special case of two-phase sampling as described in Section 1. If the selection within clusters is independent of which clusters have been selected in the first phase, then we are in the presence of a strongly invariant two-stage cluster sampling design. This is satisfied if the selection of elements within clusters is independent of the selection of elements in any other cluster. A strongly invariant two-stage cluster sampling designs can be implemented by reversing the actual act of sampling: instead of sampling the clusters first, we begin by selecting the elements in each of the population clusters, and then sampling the clusters.*

Note that our definition of strong invariance for two-stage designs is slightly different from the one given in Särndal, Swensson and Wretman (1992, Chapter 4) because the latter restrict to clusters selected at the

first stage. However, for practical purposes, both definitions are essentially equivalent. We used Definition 1 rather the standard definition of Särndal et al. (1992) because the latter does not extend easily to the case of two-phase sampling.

Definition 2. *A two-phase sampling design is said to be weakly (or first-two-moment) invariant if*

$$\pi_{2i}(\mathbf{I}_1) = \pi_{2i} \quad \text{and} \quad \pi_{2ij}(\mathbf{I}_1) = \pi_{2ij} \quad i \in s_1, j \in s_1.$$

Clearly, a strongly invariant two-phase sampling design is weakly invariant but the opposite is not true. The next example describes a sampling design that is weakly invariant but not strongly invariant.

Example 3. *At the first phase, we select a sample, s_1, of size n_1, according to an arbitrary fixed-size sampling design. From s_1, we select a simple random sample without replacement, s_2, of size n_2, where n_2 is fixed prior to sampling. This two-phase sampling design is weakly invariant since $\pi_{2i} = n_2/n_1$, and $\pi_{2ij} = n_2(n_2-1)/n_1(n_1-1)$, which remain the same from one realization of $\mathbf{I}_1$ to another. However, it is not strongly invariant since it is not possible to generate $\mathbf{I}_2$ prior to $\mathbf{I}_1$ and meet the fixed-size sample size constraint for n_2. In fact, this would also be true for any fixed-size sampling design at the second phase satisfying $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ and $\pi_{2ij}(\mathbf{I}_1) = \pi_{2ij}$.*

Finally, we describe a non-invariant two-phase sampling design.

Example 4. *At the first phase, we select a simple random sample without replacement, s_1, of size n_1, according to an arbitrary fixed-size sampling design. For every $i \in s_1$, we record an auxiliary variable x. From s_1, a second-phase sample, s_2, of fixed size n_2, is selected using an inclusion probability proportional-to-size procedure. In this case, we have*

$$\pi_{2i}(\mathbf{I}_1) = \frac{n_2 x_i}{\sum_{i \in U} x_i I_{1i}}.$$

Clearly, the inclusion probability of unit i in s_2 vary from one realization of $\mathbf{I}_1$ to another. Since $\pi_{2i}(\mathbf{I}_1)$ is a function of $\mathbf{I}_1$, it is known only after the first-phase sample s_1 is actually realized.

3 Implications of the invariance property

3.1 Weak invariance

For an arbitrary two-phase sampling design, the inclusion probability of unit i, $\pi_i, i \in s_1$, is generally unknown and is defined as

$$\begin{aligned} \pi_i &= \mathrm{E}(I_{1i} I_{2i}) \\ &= \mathrm{E}\{I_{1i}\mathrm{E}(I_{2i} \mid \mathbf{I_1})\} \\ &= \sum_{\mathbf{i}_1 : i_{1i}=1} \pi_{2i}(\mathbf{I_1}) P(\mathbf{I}_1 = \mathbf{i}_1), \end{aligned} \tag{3.1}$$

where $\mathbf{i}_1$ denotes a realisation of the random vector $\mathbf{I}_1$. Therefore, the π_i's are generally unknown because they require the knowledge of $P(\mathbf{I}_1 = \mathbf{i}_1)$ for every possible $\mathbf{I}_1$ (in many cases, we do) but also of $\pi_{2i}(\mathbf{I}_1)$ for every $\mathbf{I}_1$. The latter are generally unknown because $\pi_{2i}(\mathbf{I}_1)$ may depend on the outcome of phase 1. However, if the sampling design is weakly invariant, then $\pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ and (3.1) reduces to

$$\pi_i = \pi_{2i} \sum_{\mathbf{i}_1:\, i_{1i}=1} P(\mathbf{I}_1 = \mathbf{i}_1) = \pi_{1i}\pi_{2i}. \tag{3.2}$$

Suppose that we are interested in estimating the population total $t_y = \sum_{i \in U} y_i$. Since the π_i's are generally unknown, the Horvitz-Thompson estimator of t_y,

$$\hat{t}_{HT} = \sum_{i \in s_2} \pi_i^{-1} y_i,$$

cannot be used, in general. Instead, it is common practice to use the double expansion estimator

$$\hat{t}_{DE} = \sum_{i \in s_2} \pi_{1i}^{-1} \pi_{2i}(\mathbf{I}_1)^{-1} y_i.$$

In general, both $\hat{t}_{HT}$ and $\hat{t}_{DE}$ differ. However, for weakly invariant two-phase designs, it is clear from (3.2), that both are identical.

3.2 Strong invariance

Let θ be a finite population parameter and $\hat{\theta}$ be an estimator of θ. The total variance of $\hat{\theta}$ can be expressed as

$$V(\hat{\theta}) = VE(\hat{\theta} \mid \mathbf{I}_1) + EV(\hat{\theta} \mid \mathbf{I}_1). \tag{3.3}$$

Decomposition (3.3) is often called the two-phase decomposition of the variance; e.g., Särndal et al. (1992). If the two-phase sampling design is strongly invariant, the total variance of $\hat{\theta}$ can alternatively be decomposed as

$$V(\hat{\theta}) = EV(\hat{\theta} \mid \mathbf{I}_2) + VE(\hat{\theta} \mid \mathbf{I}_2). \tag{3.4}$$

The decomposition (3.4) is often called the reverse decomposition of the variance as the order of sampling is reversed, which can only be justified provided the two-phase design is strongly invariant. The decomposition (3.4) cannot be used in the case of weakly invariant two-phase design as the vector $\mathbf{I}_2$ cannot be generated prior to the vector $\mathbf{I}_1$. The reverse decomposition was studied in the context of nonresponse by Fay (1991), Shao and Steel (1999) and Kim and Rao (2009), among others. In a nonresponse context, assuming that the units respond independently of one another, the set of respondents can be viewed as a second-phase sample selected according to Poisson sampling with unknown inclusion probabilities, called response probabilities. If the latter remain the same from one realization of the sample to another, we are essentially in the presence of a strongly invariant two-phase sampling design. Decomposition (3.4) can be

used to justify simplified variance estimators for two-phase sampling designs; see Beaumont, Béliveau and Haziza (2015).

Acknowledgements

The authors are grateful to an Associate Editor and a reviewer for their comments and suggestions, which improved the quality of this paper. David Haziza's research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F., Béliveau, A. and Haziza, D. (2015). Clarifying some aspects of variance estimation in two-phase sampling. *Journal of Survey Statistics and Methodology*, 3, 524-542.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference,* US Bureau of the Census, 429-440.

Kim, J.K., and Rao, J.N.K. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

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

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

CORRIGENDUM

Statistical matching using fractional imputation

Jae Kwang Kim, Emily Berg and Taesung Park
Volume 42, number 1, (June 2016), 19-40

On page 21 of the original release of Kim, Berg and Park (2016), in remark (2.1), assumption (i) should be

$$(i) \quad f(y_2 | x_1, x_2, y_1) = f(y_2 | x_2, y_1)$$

That is, on the right hand side, the subscript of x should be 2 rather than 1.

Corrected electronic versions of the original paper have been uploaded.

Reference

Kim, J.K., Berg, E. and Park, T. (2016). Statistical matching using fractional imputation. *Survey Methodology*, 42, 1, 19-40.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2016.

P. Aronow, *Yale University*

J.-F. Beaumont, *Statistics Canada*

B. Bell, *U.S. Census Bureau*

M. Berzofsky, *Research Triangle Institute International*

D. Bonnery, *University of Maryland*

C. Boulet, *Statistics Canada*

J. Breidt, *Colorado State University*

J. Brown, *School of Mathematical and Physical Sciences*

S. Casanova, *Université de Toulouse*

J. Chen, *University of British Columbia*

R. Clark, *University of Wollongong*

M.H. Felix-Medina, *Universidad Autonoma de Sinaloa*

S. Fortier, *Statistics Canada*

C. Girard, *Statistics Canada*

J.M. Gonzalez, *U.S. Bureau of Labor Statistics*

D. Haziza, *Université de Montréal*

M.Y. Hirose, *Institute of Statistical Mathematics*

G. Kalton, *Westat Inc.*

C. Kennedy, *Pew Research Center*

N. Kim, *Carnegie Mellon University*

O. Kitov, *Oxford University*

P. Kott, *Research Triangle Institute International*

P. Lahiri, *University of Maryland*

A. Lauger, *U.S Census Bureau*

H. Lee, *Westat Inc.*

C. Leger, *Université de Montréal*

B. Levine, *Research Triangle Institute International*

J. Lim, *Seoul National University*

P. Linde, *Statistics Denmark*

S. Lohr, *Westat Inc.*

H. Mantel, *Statistics Canada*

A. Matei, *University of Neuchatel*

G.H. McLaren, *Office for National Statistics*

I. McLeod, *University of Western Ontario*

T. Merkouris, *Athens University of Economics and Business*

J.F. Munoz Rosas, *University of Granada*

J. Opsomer, *Colorado State University*

G. Osier, *STATEC Luxembourg*

S. Pal, *West Bengal State University*

D. Pfeffermann, *Southampton Statistical Sciences Research Institute*

W. Qian, *Statistics Canada*

M.G. Ranalli, *Dept. of Political Science, University of Perugia*

J. Reiter, *Duke University*

A. Ruiz-Gazen, *Toulouse School of Economics*

N. Salvati, *University of Pisa*

B. Schouten, *Statistics Netherlands and Utrecht University*

P.A. Smith, *University of Southampton*

R. Tiller, *Bureau of Labour Statistics*

X. Wang, *Southern Methodist University*

D. Wilson, *Research Triangle Institute International*

C. Wu, *University of Waterloo*

Acknowledgements are also due to those who assisted during the production of the 2016 issues: Céline Ethier of International Cooperation and Corporate Statistical Methods Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Christina Jaworski, Kathy Charbonneau, Giovanni Borrello, Joseph Prince et Darquise Pellerin as well as our partners in the Communications Division.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 32, No. 2, 2016

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 32, No. 3, 2016

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 44, No. 3, September/septembre 2016

Cover

Issue Information

Original Articles

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 44, No. 4, December/décembre 2016

Cover

Issue Information

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The documents should be divided into numbered sections with suitable verbal titles.
1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.
4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.