

A NOTE ON SAMPLING AND ESTIMATION IN THE PRESENCE OF CUT-OFF SAMPLING

David Haziza, Guillaume Chauvet and Jean-Claude Deville¹

October 7, 2009

ABSTRACT

Cut-off sampling consists of deliberately excluding a set of units from possible selection in sample, for example if the contribution of the excluded units to the total is small or if the inclusion of these units in the sample involves high costs. If the characteristics of interest of the excluded units differ from that of the rest of the population, the use of naïve estimators may result in highly biased estimates. In this paper, we discuss the use of auxiliary information to reduce the bias by means of calibration and balanced sampling. We show that the use of the available auxiliary information related to both the variable of interest and the probability of being excluded enables us to reduce the potential bias. A short numerical study supports our findings.

KEYWORDS: Auxiliary information; balanced sampling; calibration; cut-off sampling; design bias; model bias.

¹ David Haziza (David.Haziza@umontreal.ca), Département de mathématiques et de statistique, Université de Montréal, Montréal, Québec, H3C 3J7, Canada. Guillaume Chauvet (chauvet@ensai.fr) and Jean-Claude Deville (deville@ensai.fr), Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France.

1. INTRODUCTION

Cut-off sampling, which consists of deliberately excluding a set of units from possible selection in sample, is frequently used in business surveys where the small businesses are often grouped into take-none strata. The contribution to the overall total of the excluded units is typically small. The main reason motivating the use of cut-off sampling is that even though the resulting estimators generally suffer from a slight loss in accuracy, the survey costs may be significantly reduced.

Another example of cut-off sampling occurs in the context of tax data for unincorporated businesses at Statistics Canada. The unincorporated Canadian businesses may declare their financial statement either on paper or electronically (e.g., using internet). The businesses that use a paper format are called the paper-filers or p-filers, whereas the ones that choose the electronic format are called the electronic-filers or e-filers. A little bit more than half of the businesses (52%) belongs to the population of *e*-filers (see Fecteau and Jocelyn, 2005). The population under study, U , of size N , can thus be partitioned into two strata: the strata of *e*-filers, U_E , of size N_E , and the strata of *p*-filers, U_P , of size N_P . We have $U = U_E \cup U_P$ and $N = N_E + N_P$. Fecteau and Jocelyn (2005) mention that the populations of *e*-filers and *p*-filers have different characteristics. Typically, the businesses using a paper format are larger in size than the ones using an electronic format and thus have a larger income (see Table 1).

The goal is to produce estimates for totals in U (e.g., gross and net income) based on a sample of businesses. However, due to high costs of converting data collected on paper to an electronic format, the *p*-filers are deliberately excluded from a possible selection in the sample and the resulting estimates are based on a sample selected in the strata of *e*-filers only.

Thus, this sampling procedure can be viewed as a special case of cut-off sampling. The main concern is that, since the p-filers have a sample inclusion probability equal to 0, we cannot construct design-unbiased estimators of population totals (or means). Also, the use of naïve estimators may potentially lead to severely biased estimators of population totals since both populations (e-filers and p-filers) differ with respect to several characteristics of interest. The key to reduce the bias due to the exclusion of the p-filers from the sample is to use the available auxiliary information.

In this paper, we consider two well known techniques that incorporate some amount of auxiliary information: (i) balanced sampling and (ii) calibration. As we argue in sections 2 and 3, both techniques can help reduce the bias due to cut-off sampling provided powerful auxiliary information is available. Unlike calibration, balanced sampling requires the auxiliary information to be available for all the units in the population. Although the results presented in this paper are derived in the context of the e-filers and p-filers, they can be applied to any type of cut-off sampling conducted under the same conditions.

We are interested in estimating the population total, $Y = \sum_{i \in U} y_i$, of a given variable of interest

y . Note that Y may be expressed as $Y = Y_E + Y_P$, where $Y_E = \sum_{i \in U_E} y_i$ and $Y_P = \sum_{i \in U_P} y_i$. To that

end, we select a random sample, s_E , of size n_E , according to a given design $p_E(\cdot)$ from U_E .

Let $d_i = 1/\pi_i$ be the design weight attached to unit i , where $\pi_i = P(i \in s_E)$ is the first-order inclusion probability of unit i in the sample s_E . We assume that the variable y is observed for

all $i \in s_E$. The main issue is that since $\pi_i = 0$ for all $i \in U_P$, a design-unbiased estimator of

Y does not exist.

A basic estimator of Y is the so-called Hajek estimator (Hajek, 1971) given by

$$\hat{Y}_{HA} = N\bar{y}_E, \quad (1.1)$$

where $\bar{y}_E = \frac{\hat{Y}_E}{\hat{N}_E}$ with $(\hat{Y}_E, \hat{N}_E) = \sum_{i \in S_E} d_i(y_i, 1)$.

In order to evaluate the properties of point estimators, we consider a number of approaches for inference. To understand the nature of the different approaches, we first identify three sources of randomness: (i) the superpopulation model m , which generates the vector of population values $\mathbf{y} = (y_1, \dots, y_N)'$; (ii) the e-filer/p-filer mechanism q which generates the vector of e-filer/p-filer indicators $\mathbf{a} = (a_1, \dots, a_N)'$ such that $a_i = 1$ if unit i is an e-filer and $a_i = 0$, otherwise. The variable a_i is assumed to be known for all $i \in U$; (iii) the sampling design $p(\cdot)$, which generates the vector of sample indicators $\mathbf{I} = (I_1, \dots, I_N)'$ such that $I_i = 1$ if unit i is selected in the sample and $I_i = 0$, otherwise. Based on these random vectors, there are several possible approaches to inference. The first is the s -approach which treats the vectors \mathbf{y} and \mathbf{a} as fixed. Under this approach, the only remaining source of randomness is the vector of sample selection indicators \mathbf{I} . When an estimator is unbiased under the s -approach, we say that the estimator is s -unbiased. The second approach is the sm -approach, under which the properties of estimators are evaluated with respect to the joint distribution induced by the superpopulation model and the sampling design, whereas the vector \mathbf{a} is treated as fixed. When an estimator is unbiased under the sm -approach, we say that the estimator is sm -unbiased. Finally, the third approach is the sq -approach, which consists of evaluating the properties of estimators with respect to the joint distribution induced by the e-filer/p-filer mechanism and the sampling design, and treating the vector as fixed. When an estimator is unbiased under the sq -approach, we say that the estimator is sq -unbiased.

Let $E_s(\cdot)$ denote the expectation with respect to the sampling design. The s -bias of \hat{Y}_{HA} ,

$B_s(\hat{Y}_{HA}) = E_s(\hat{Y}_{HA}) - Y$, can be approximated by

$$B_s(\hat{Y}_{HA}) \approx N_p(\bar{Y}_E - \bar{Y}_p), \quad (1.2)$$

where $\bar{Y}_E = \frac{Y_E}{N_E}$ and $\bar{Y}_p = \frac{Y_p}{N_p}$ denote the population mean of the y -values for the e-filers and

p-filers, respectively. The bias in (1.2) is equal to 0 if $N_p = 0$, which occurs when $U_p = \emptyset$, or

when $\bar{Y}_E = \bar{Y}_p$. These two conditions are not satisfied in practice since the population of p-

filers represents approximately 48% of the population and since the two populations differ

with respect to several characteristics such as *Gross income* (see Table 1). Therefore, the

Hajek estimator \hat{Y}_{HA} may be severely s -biased when \bar{Y}_E is significantly different than \bar{Y}_p .

Note that we have $\bar{Y}_E = \bar{Y}_p$ when the variable of interest y is not related to the e-filer/p-filer

status. Alternative strategies are thus needed and are presented in sections 2 and 3.

The paper is organized as follows: in section 2, we present three classes of calibration

estimators and study their properties in terms of bias. In section 3, we present two versions of

balanced sampling and discuss the properties of the resulting estimators. A limited simulation

study is conducted in section 4 in order to investigate the performance of the proposed

estimators in terms of bias and mean square error. Finally, we conclude in section 5.

Table 1: Mean income of businesses by type of format

Format	Gross mean income	Net mean income
Electronic	261 819 \$	11 712 \$
Paper	694 587 \$	14 021 \$

2. CALIBRATION ESTIMATORS

In this section, we study the use of auxiliary information through calibration for reducing the bias.

Suppose that a vector of J auxiliary variables $\mathbf{x} = (x_1, \dots, x_J)'$ is available for all the units in the sample s_E and that the vector of population totals, $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$, is known. We assume that the relationship between the variable of interest y and the vector of auxiliary variables \mathbf{x} is described according to the following model:

$$m: y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad (2.1)$$

such that $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$ if $i \neq j$ and $V_m(\varepsilon_i) = \sigma^2 c_i$, where $\boldsymbol{\beta}$ is a J -vector of unknown parameters, σ^2 is an unknown parameter, c_i is a known constant and $E_m(\cdot)$ and $V_m(\cdot)$ denote the expectation and the variance with respect to the model (2.1), respectively.

We assume that $c_i = \boldsymbol{\alpha}' \mathbf{x}_i$, where $\boldsymbol{\alpha}$ is a J -vector of specified constants.

2.1 Direct calibration

A first set of estimators can be obtained via direct calibration which consists of finding a set of new weights, $w_i = d_i F(c_i^{-1} \boldsymbol{\lambda}' \mathbf{x}_i)$, so that the calibration equations

$$\sum_{i \in s_E} w_i \mathbf{x}_i = \mathbf{X} \quad (2.2)$$

are satisfied, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers and $F(\cdot)$ is the so-called calibration function. Several calibration functions $F(\cdot)$ can be used; see Deville and Särndal (1992), Deville (2002), Le Guennec and Sautory (2002) and Kott (2006), among others. Two such

options for $F(\cdot)$ are : (i) the linear function, $F(u)=1+u$, which corresponds to the generalized chi-square distance and (ii) the exponential function, $F(u)=e^u$, which corresponds to the raking ratio distance; see Deville and Särndal (1992). Except for the linear case for which we can obtain an explicit solution, the Newton-Raphson algorithm is needed for solving (2.2) for general $F(\cdot)$. The resulting calibration estimator of Y is given by

$$\hat{Y}_{CAL} = \sum_{i \in s_E} d_i F(c_i^{-1} \boldsymbol{\lambda}' \mathbf{x}_i) y_i. \quad (2.3)$$

When $F(u)=1+u$ in (2.3), we obtain the generalized regression estimator

$$\hat{Y}_G = \sum_{i \in s_E} w_i y_i, \quad (2.4)$$

where

$$w_i = d_i \left[1 + c_i^{-1} (\mathbf{X} - \hat{\mathbf{X}}_E)' \hat{\mathbf{T}}_E^{-1} \mathbf{x}_i \right] \quad (2.5)$$

with $\hat{\mathbf{T}}_E = \sum_{i \in s_E} d_i c_i^{-1} \mathbf{x}_i \mathbf{x}_i'$ and $\hat{\mathbf{X}}_E = \sum_{i \in s_E} d_i \mathbf{x}_i$. Note that the Hajek estimator given by (1.1) is a special case of (2.4) with $\mathbf{x}_i = 1$ and $c_i = 1$.

The asymptotic s -bias of \hat{Y}_G in (2.4), $B_s(\hat{Y}_G) = E_s(\hat{Y}_G) - Y$, is given by

$$\begin{aligned} B_s(\hat{Y}_G) &\approx - \sum_{i \in U_p} (y_i - \mathbf{x}_i' \mathbf{B}_E) \\ &= \mathbf{X}_P' (\mathbf{B}_E - \mathbf{B}_P), \end{aligned} \quad (2.6)$$

where $\mathbf{X}_P = \sum_{i \in U_p} \mathbf{x}_i$, $\mathbf{B}_E = \left[\sum_{i \in U_E} c_i^{-1} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i \in U_E} c_i^{-1} \mathbf{x}_i y_i$ denotes the census coefficient of regression corresponding to the population of e-filers and \mathbf{B}_P is defined similarly. The asymptotic s -bias of \hat{Y}_G in (2.6) is small if $\sum_{i \in U_p} (y_i - \mathbf{x}_i' \mathbf{B}_E)$ is small, which occurs, for

example, when the residuals $E_i = (y_i - \mathbf{x}'_i \mathbf{B}_E)$ corresponding to the p-filers are small, which in turn suggests that the model (2.1) holds for the p-filers. Also, if the e-filer/p-filer status is not related to the variable of interest y after accounting for \mathbf{x} , we expect to have $\mathbf{B}_E \approx \mathbf{B}_P$ and the s -bias of \hat{Y}_G is asymptotically equal to 0. This would occur, for example, if the probability p_i for unit i to be an e-filer is constant for all i .

On the other hand, if the model (2.1) holds (i.e., the model holds for both the p-filers and the e-filers), the estimator \hat{Y}_G is sm -unbiased. That is, $B_{sm}(\hat{Y}_G) = E_s E_m(\hat{Y}_G - Y) = 0$. However, since no y -value are observed for the p-filers, validating the model may prove to be difficult. To illustrate the difficulty, we consider the following example. Suppose that the model that holds for the e-filers is obtained from (2.1) by replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_E$, whereas the model that holds for the p-filers is obtained from (2.1) by replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_P$. In this case, the estimator \hat{Y}_G is asymptotically sm -biased and the asymptotic bias is given by

$$B_{sm}(\hat{Y}_G) \approx \mathbf{X}'_P (\boldsymbol{\beta}_E - \boldsymbol{\beta}_P). \quad (2.7)$$

From (2.7), it is clear that the bias of \hat{Y}_G depends on the difference of $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_P$. From the observed data, it is not possible to assess the magnitude of this difference since no y -values are available for the p-filers. This example illustrates the problem of building an appropriate model. As a result, reducing the bias may prove to be difficult. In fact, if $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_P$ are considerably different, the bias of \hat{Y}_G may be even larger than the bias of the naïve estimator, \hat{Y}_{HA} , given by (1.1).

2.2 Calibration after reweighting

As discussed in section 2.1, the regression estimator (2.4) may present some risks when it is not possible to validate the model at hand. In this section, we propose an alternative class of calibration estimators that may be more robust to bias. In addition to the vector of auxiliary variables \mathbf{x} related to the variable of interest y , we assume that there exists a l -vector of auxiliary variables \mathbf{z}_i related to the probability of e-filer/p-filer status, $p_i = P(a_i = 1)$. We assume that the vector \mathbf{z} is available for all the units in the population. Note that $\mathbf{x}_i \neq \mathbf{z}_i$, in general. The relationship between p_i and \mathbf{z}_i may be described according to the following parametric model:

$$\xi: p_i = f(\mathbf{z}_i, \boldsymbol{\gamma}), \quad (2.8)$$

where $f(\cdot)$ is a given function, and $\boldsymbol{\gamma}$ is a l -vector of unknown parameters. A special case of (2.8), which is frequently used in practice, is the logistic regression model given by

$$p_i = \frac{e^{\mathbf{z}_i' \boldsymbol{\gamma}}}{1 + e^{\mathbf{z}_i' \boldsymbol{\gamma}}}. \quad (2.9)$$

Let \hat{p}_i be the estimated probability for unit i given by

$$\hat{p}_i = f(\mathbf{z}_i, \hat{\boldsymbol{\gamma}}),$$

where $\hat{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma}$ (usually the maximum likelihood estimator of $\boldsymbol{\gamma}$).

Alternatively, the estimated probabilities \hat{p}_i may be obtained using a nonparametric method such as kernel-type smoothing methods or local polynomial regression; e.g., Wand and Jones (1995).

A second class of estimators can be obtained by finding a new set of weights

$w_i^* = d_i^* F(c_i^{-1} \boldsymbol{\lambda}' \mathbf{x}_i)$, so that the calibration equations

$$\sum_{i \in S_E} w_i^* \mathbf{x}_i = \mathbf{X} \quad (2.10)$$

are satisfied, where $d_i^* = d_i / \hat{p}_i$. The resulting calibration estimator of Y is given by

$$\hat{Y}_{CAL}^* = \sum_{i \in s_E} d_i^* F(c_i^{-1} \boldsymbol{\lambda}' \mathbf{x}_i) y_i. \quad (2.11)$$

In some cases, direct calibration, presented in section 2.1, and calibration after reweighting lead to identical estimators. For example, if the model ξ given by (2.8) contains only the intercept (i.e., $\mathbf{z}_i = 1$), then the alternative calibration estimator (2.11) reduces to the estimator (2.3). Also, if the estimated response probabilities are obtained by fitting the logistic model (2.9) with $\mathbf{z}_i = \mathbf{x}_i$ and $F(u) = e^u$, then the estimator (2.11) is identical to the estimator (2.3) that uses $F(u) = e^u$.

In the special case of the linear function, $F(u) = 1 + u$, the estimator (2.11) reduces to

$$\hat{Y}_G^* = \sum_{i \in s_E} w_i^* y_i, \quad (2.12)$$

where $w_i^* = d_i^* \left(1 + c_i^{-1} (\mathbf{X} - \hat{\mathbf{X}}_E^*)' \hat{\boldsymbol{\Gamma}}_E^{*-1} \mathbf{x}_i \right)$ with $\hat{\boldsymbol{\Gamma}}_E^* = \sum_{i \in s_E} d_i^* c_i^{-1} \mathbf{x}_i \mathbf{x}_i'$ and $(\hat{\mathbf{X}}_E^*, \hat{Y}_E^*) = \sum_{i \in s_E} d_i^* (\mathbf{x}_i, y_i)$.

If the assumed model (2.8) is valid (i.e., \hat{p}_i is q -consistent for p_i), then the alternative regression estimator \hat{Y}_G^* in (2.12) is asymptotically sq -unbiased for Y . That is, $B_{sq}(\hat{Y}_G^*) = E_s E_q(\hat{Y}_G^*) \approx Y$. This property holds even if the model (2.1) does not hold. On the other hand, if the model (2.1) holds for the e-filers and p-filers, the estimator \hat{Y}_G^* is asymptotically sm -unbiased for Y . That is, $B_{sm}(\hat{Y}_G^*) = E_s E_m(\hat{Y}_G^* - Y) \approx 0$. This property holds even if the model (2.8) is misspecified. Hence, \hat{Y}_G^* is doubly robust in the sense that it is valid if one model or the other (m or ξ) holds.

Unlike in the case of model (2.1), the model (2.8) can easily be validated from the values of a_i and \mathbf{z}_i for $i \in U$ since the indicator variable a_i is available for all the units in the population. Note that the estimator \hat{Y}_G^* given by (2.12) uses the available auxiliary information (\mathbf{x} and \mathbf{z}) to a greater extent than \hat{Y}_G , which only uses \mathbf{x} . As a result, it is expected that the use of \hat{Y}_G^* will achieve an effective bias reduction if either model (m or ξ) holds. Another advantage of \hat{Y}_G^* over \hat{Y}_G is that for surveys with multiple characteristics, the estimator \hat{Y}_G^* is asymptotically sq -unbiased for any variable of interest y , as long as the probabilities \hat{p}_i are correctly estimated. On the other hand, the estimator \hat{Y}_G may be asymptotically sm -unbiased for the total of a given variable of interest but not necessarily asymptotically sm -unbiased for the total of another variable of interest, as the set of auxiliary variables that explain the two variables may not be identical.

It is worth noting that the vector \mathbf{z} should include the variables that are related to both the probability p_i as well as the variable of interest. If an auxiliary variable is related to the e-filer/p-filer status but not to the variables of interest, it should not be used since it will not help reducing the bias but is likely to contribute in increasing the variance of the estimators. In the context of the e-filers and p-filers, the goal is to estimate the total of the variable *Gross income* within industries. An example of an x variable is *Net income*, whereas the variable *Business Size* (e.g., number of employees) could be used as a z variable since it is related to both the e-filers/p-filers status and *Gross income*.

In practice, the partition of the population into a take-none portion and a take-some portion is often performed in a deterministic fashion. For example, the businesses whose revenue r_i

is smaller than a given threshold r_0 are excluded from possible selection in sample. In this case, we have $p_i = \hat{p}_i = 1$ if unit i is an e-filer and $p_i = \hat{p}_i = 0$, otherwise. As a result, the estimator (2.12) is biased since $p_i = 0$ if i is a p-filer. To overcome this problem, one can obtain estimated respond probabilities using z-variables available at the estimation stage, excluding the variable revenue that was used to partition the population. That is, one may use a parametric model

$$p(r_i \leq r_0) = f(z_i, \gamma)$$

analog to the model (2.8), where the probability for a business to have a revenue smaller than the threshold r_0 is modeled in terms of the z-variables available at the estimation stage.

2.3 Generalized Calibration

A second class of estimators that makes use of the vectors of auxiliary variables \mathbf{x} and \mathbf{z} , is obtained by generalized calibration presented in Deville (1998; 2002), Sautory (2003) and Kott (2006), among others. It assumes that the vectors \mathbf{x} and \mathbf{z} are of the same dimension. One advantage of generalized calibration over the calibration methods presented in section 2.1 and 2.2 is that the \mathbf{z} -values are required for the e-filers only. In fact, the vector \mathbf{z} may include the variable of interest y as one component, which may help in reducing the bias to a greater extent. Here, we seek a new set of weights $\tilde{w}_i = d_i F(\lambda' \mathbf{z}_i)$ so that the calibration equations

$$\sum_{i \in s_E} \tilde{w}_i \mathbf{x}_i = \mathbf{X} \quad (2.13)$$

are satisfied. The resulting calibration estimator of Y is given by

$$\hat{Y}_{CAL} = \sum_{i \in s_E} d_i F(\lambda' \mathbf{z}_i) y_i. \quad (2.14)$$

In the special case of the linear function, $F(u) = 1 + u$, the estimator (2.14) reduces to

$$\hat{Y}_G = \sum_{i \in S_E} \tilde{w}_i y_i, \quad (2.15)$$

where $\tilde{w}_i = d_i \left(1 + (\mathbf{X} - \hat{\mathbf{X}}_E)' \hat{\mathbf{T}}_E^{-1} \mathbf{z}_i \right)$ with $\hat{\mathbf{T}}_E = \sum_{i \in S_E} d_i \mathbf{z}_i \mathbf{x}_i'$. Note that the estimator (2.15) can

be expressed as

$$\hat{Y}_G = \hat{Y}_E + (\mathbf{X} - \hat{\mathbf{X}}_E)' \hat{\mathbf{B}}_E, \quad (2.16)$$

where $\hat{\mathbf{B}}_E = \hat{\mathbf{T}}_E^{-1} \hat{\mathbf{t}}_E$ with $\hat{\mathbf{t}}_E = \sum_{i \in S_E} d_i \mathbf{z}_i y_i$. Thus, the estimated regression coefficient $\hat{\mathbf{B}}_E$ can be

viewed as the estimated regression coefficient obtained by fitting an instrumental regression with the vector \mathbf{z} as the instrument. If the vector \mathbf{z} explains well the e-filer/p-filer status, the estimator (2.16) is expected to have a small bias. The vectors \mathbf{x} and \mathbf{z} have to be strongly correlated. Otherwise, the matrix $\tilde{\mathbf{T}}_E = \sum_{i \in U_E} \mathbf{z}_i \mathbf{x}_i'$ may be close to singularity, which may result

in a highly variable calibrated estimator \hat{Y}_{CAL} . Note that when $\mathbf{z}_i = c_i^{-1} \mathbf{x}_i$, the estimator (2.14) reduces to the direct calibration estimator given by (2.3).

Under the sq -approach, the asymptotic bias of \hat{Y}_G , $B_{sq}(\hat{Y}_G) = E_s E_q(\hat{Y}_G) - Y$, can be approximated by

$$B_{sq}(\hat{Y}_G) \approx - \sum_{i \in U} (1 - p_i) (y_i - \mathbf{x}_i' \tilde{\mathbf{B}}_p), \quad (2.17)$$

where $\tilde{\mathbf{B}}_p = \left[\sum_{i \in U} p_i \mathbf{z}_i \mathbf{x}_i' \right]^{-1} \sum_{i \in U} p_i \mathbf{z}_i y_i$. The asymptotic bias in (2.17) is equal to zero if the probability p_i is not related to \mathbf{x} after accounting for \mathbf{z} . However, if there remains a residual relationship between p_i and \mathbf{x} after accounting for \mathbf{z} , the estimator \hat{Y}_G may be severely biased. This is illustrated in the simulation study described in section 4.

On the other hand, the estimator \hat{Y}_G is asymptotically *sm*-unbiased provided the model (2.1) holds for both the e-filers and the p-filers.

3. BALANCED SAMPLING

A balanced sampling design ensures that estimators of the auxiliary variables, called balancing variables, match the known totals exactly. The estimator of a variable of interest y is still design-unbiased, while its variance is strongly reduced if the balancing variables are highly related to y . Deville and Tillé (2004) proposed a general algorithm for balanced sampling with any set of unequal probabilities and a non-restricted number of balancing variables.

The use of auxiliary information through balanced sampling for reducing the bias is discussed in the following sections. In section 3.1, a method for directly selecting a balanced sample is given. In sections 3.2 and section 3.3, alternative balanced sampling strategies to reduce the estimation bias are proposed. We assume that a J -vector of balancing variables, $\mathbf{x} = (x_1, \dots, x_J)'$, is known for all the units of the population.

3.1 The traditional balanced sampling

In this section, we briefly introduce the Cube method (see Deville and Tillé, 2004). First, we shall assume that the stratum U_p is empty; that is, that the sample is selected from U such that $\pi_i > 0$ for all i , and $s_E = s$. Recall that $\hat{\mathbf{X}} = \sum_{i \in s} d_i \mathbf{x}_i$ is a s -unbiased estimator of

$$\mathbf{X} = \sum_{i \in U} \mathbf{x}_i; \text{ that is,}$$

$$E_s(\hat{\mathbf{X}}) = \mathbf{X}. \quad (3.1)$$

A sampling design is said to be balanced on the variables \mathbf{x} if the balancing equations

$$\hat{\mathbf{X}} = \mathbf{X} \quad (3.2)$$

are satisfied. Note that (3.2) holds exactly unlike (3.1) that holds on average. In the special case of $\mathbf{x}_i = \pi_i$, $\hat{\mathbf{X}}$ and \mathbf{X} represent the actual sample size and the expected sample size respectively, and condition (3.2) is equivalent to imposing a fixed sample size. When $\mathbf{x}_i = 1$ for all i , $\hat{\mathbf{X}}$ and \mathbf{X} represent the estimated and true population size respectively and expression (3.2) means that the sample s is such that the estimated population size matches the true population size.

The Cube method proposed by Deville and Tillé (2004) provides a general algorithm for selecting balanced samples with predetermined inclusion probabilities. An exact balanced sampling design generally does not exist, that is, there may not exist any sample such that equation (3.2) holds. Consequently, the objective is to select a sample such that the balancing equations (3.2) hold approximately in the sense that

$$\hat{\mathbf{X}} = \mathbf{X} + O\left(\frac{N \times J}{n}\right),$$

whereas we have

$$\hat{\mathbf{X}} = \mathbf{X} + O_p\left(\frac{N}{\sqrt{n}}\right)$$

if the sampling design is not balanced on \mathbf{x} , under a standard asymptotic framework for finite population sampling with $N \rightarrow \infty$. The Cube algorithm is thus divided into two steps. In the first step, units are sampled or definitely rejected so that both the inclusion probabilities and the balancing equations are exactly satisfied. This step stops when the balancing conditions can no longer be exactly satisfied. The last step consists in ending the sampling so that the

inclusion probabilities remain exactly satisfied and the balancing conditions remain approximately satisfied. In the context of balanced sampling, note that the vector of balancing variables \mathbf{x} must be known prior to sampling for all the population units, unlike in the calibration context for which the vector \mathbf{x} is required for the sample units and only the vector of totals \mathbf{X} must be known.

3.2 The corrected balanced sampling

We now turn back to the general setting of non empty stratum U_p . We assume that the relationship between the variable of interest y and the vector of auxiliary variables \mathbf{x} may be described according to the model (2.1). Suppose that the sample s_E is selected from U_E by balanced sampling with inclusion probability π_i for unit $i \in U_E$ and balancing variables \mathbf{x} , so that the equations

$$\sum_{i \in s_E} d_i \mathbf{x}_i = \sum_{i \in U_E} \mathbf{x}_i \quad (3.3)$$

hold. The Hajek estimator $\hat{Y}_{HA} = N\bar{y}_E$ given by (1.1) may then be used for estimating the total Y , but this estimator remains s-biased. Alternative balanced sampling strategies are thus needed.

One such alternative consists in selecting a sample s_E by means of the Cube method, with adequate inclusion probabilities and balancing variables, so that the somewhat different balancing equations

$$\frac{1}{n} \sum_{i \in s_E} \mathbf{x}_i = \bar{\mathbf{X}} \quad (3.4)$$

are satisfied, where $\bar{\mathbf{X}} = \frac{\mathbf{X}}{N}$ denotes the overall population mean of the \mathbf{x} - vector. Condition

(3.4) may be obtained as follows. First, determine a set of weights w_i for units i of U_E such that

$$\begin{cases} \sum_{i \in U_E} w_i \mathbf{x}_i = \mathbf{X} \\ \sum_{i \in U_E} w_i = N \\ \forall i \quad w_i > 0 \end{cases}$$

These weights may be obtained by means of calibration with the raking ratio method that ensures that all calibration weights are strictly non-negative (e.g., Deville, Särndal and Sautory, 1993). Then, let n be an integer such that

$$n \frac{w_i}{N} < 1 \quad \text{for any } i \in U_E . \quad (3.5)$$

If the sample s_E is selected with inclusion probability $\pi_i = nw_i/N$ for unit i , then

$$\sum_{i \in s_E} \frac{w_i \mathbf{x}_i}{\pi_i} = \sum_{i \in s_E} \frac{w_i \mathbf{x}_i}{nw_i / N} = \frac{N}{n} \sum_{i \in s_E} \mathbf{x}_i .$$

If the sample is balanced on variables $w_i \mathbf{x}_i$, the balancing equations (3.3) lead to

$$\sum_{i \in s_E} \frac{w_i \mathbf{x}_i}{\pi_i} = \sum_{i \in U_E} w_i \mathbf{x}_i = \mathbf{X}$$

so that condition (3.4) is fulfilled. The equation (3.5) ensures that no inclusion probability exceeds 1. Selecting the higher integer n such that (3.5) holds ensures that the sample size is maximized. An estimator of Y is then given by

$$\hat{Y}_E^{un} = N \bar{y}_E^{un} \quad (3.6)$$

where $\bar{y}_E^{un} = \frac{1}{n} \sum_{i \in s_E} y_i$ is the unweighted sample mean of the y -values. The estimator \hat{Y}_E^{un} in

(3.6) is called the corrected balanced estimator, and is asymptotically m -unbiased under the model (2.1). An advantage of the corrected balanced sampling strategy is that estimator \hat{Y}_E^{un}

makes no use of the design weights d_i . On the other hand, the estimator may be seriously biased if the model (2.1) is incorrectly specified.

3.3 Balanced sampling after reweighting

In this section, we propose an estimator obtained by a modified balanced sampling design, which may be more robust to bias. Our set-up is that of section 2.2.

Once again, we assume that there exists a vector of auxiliary variables \mathbf{z}_i related to the probability $p_i = P(a_i = 1)$. Assume that p_i may be described according to the model

$$p_i^{-1} = H(\mathbf{z}'_i \boldsymbol{\gamma}) \quad (3.7)$$

for some function $H(\cdot)$. We seek estimated probabilities \hat{p}_i that satisfy the system of estimating equations

$$\sum_{i \in U_E} \frac{\mathbf{x}_i}{\hat{p}_i} = \sum_{i \in U} \mathbf{x}_i,$$

or equivalently,

$$\sum_{i \in U_E} H(\mathbf{z}'_i \hat{\boldsymbol{\gamma}}) \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i. \quad (3.8)$$

A solution for (3.8) is obtained by using the generalized calibration technique described in section 2.3. If the sample s_E is selected in U_E by means of balanced sampling with inclusion probability π_i for unit i and balancing variables \mathbf{x}/\hat{p} , an estimator of Y is given by

$$\hat{Y}_E^* = \sum_{i \in s_E} d_i^* y_i \quad (3.9)$$

with $d_i^* = d_i/\hat{p}_i$. The estimator \hat{Y}_E^* in (3.9) is called the corrected after reweighting balanced estimator. It is asymptotically sq -unbiased if the model (2.9) is valid, and also asymptotically sm -unbiased if the model (2.1) is valid. Now, suppose that

$$\mathbf{x}'_i \boldsymbol{\delta} = 1 \quad (3.10)$$

for some vector $\boldsymbol{\delta}$, and that inclusion probabilities π_i are chosen to be proportional to $1/\hat{p}_i$.

Then

$$\pi_i = n \frac{1/\hat{p}_i}{\sum_{j \in U_E} 1/\hat{p}_j}, \quad (3.11)$$

and equations (3.8) together with (3.10) lead to $\pi_i = n/(N\hat{p}_i)$. Consequently, $\hat{Y}_E^* = \frac{N}{n} \sum_{i \in s_E} y_i$

and the design weights are not needed. Note that equation (3.10) is satisfied when the vector \mathbf{x} contains the variable 1 as one of the components.

4. SIMULATION STUDY

We conducted a limited simulation study to test the performance of the procedures described in sections 2 and 3. We first generated 3 finite populations of size $N = 10\,000$, each containing three variables of interest, y_1 , y_2 and y_3 and two auxiliary variables x_1 and x_2 . First, the variables x_1 and x_2 were generated independently from a Gamma distribution with shape and scale parameters equal to 2 and 5, respectively in all the populations. In each population, the y_1 -values were generated given the x_1 -values and the x_2 -values according to the model

$$y_{1i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \eta_i. \quad (4.1)$$

The parameters β_0, β_1 and β_2 were all set to 1, whereas the η_i 's were generated according to a normal distribution with mean 0 and variance σ^2 . The parameter σ^2 was chosen to lead to a coefficient of determination (R^2) equal to 0.3 (for population 1), 0.5 (for population 2) and

0.7 (for population 3). In each population, the y_2 -values were generated given the x_1 -values and the x_2 -values according to the model

$$y_{2i} = \phi_0 + \phi_1 x_{1i} + \phi_2 x_{2i} + \phi_3 x_{2i}^2 + v_i. \quad (4.2)$$

That is, we included a quadratic term in the x_2 -values. The parameters ϕ_0, ϕ_1, ϕ_2 and ϕ_3 were all set to 1, whereas the v_i 's were generated according to a normal distribution with mean 0 and variance ζ^2 . Once again, the parameter ζ^2 was chosen to lead to a coefficient of determination (R^2) equal to 0.3 (for population 1), 0.5 (for population 2) and 0.7 for population 3. Finally, in each population, the y_3 -values were generated independently of x_1 and x_2 from a Gamma distribution with shape and scale parameters equal to 2 and 5, respectively.

Each population was partitioned into a stratum of e-filers and a stratum of p-filers, as follows: first, a probability was assigned to each population unit according to three mechanisms:

Mechanism 1: The population U was divided into four groups, U_1, \dots, U_4 , according to the quartiles of the y_1 -values. Then, a probability p_{1i} was assigned to unit i such that $p_{1i} = 0.5$ if $i \in U_1$, $p_{1i} = 0.6$ if $i \in U_2$, $p_{1i} = 0.7$ if $i \in U_3$ and $p_{1i} = 0.8$ if $i \in U_4$. Note that the average of the p_{1i} 's was equal to 65%.

Mechanism 2: The probability p_{2i} was generated for unit i according to the model

$$p_{2i} = \exp(-\gamma_0 y_{1i} / \bar{Y}_1), \quad (4.3)$$

where \bar{Y}_1 denotes the population mean of the y_1 -values. The parameter γ_0 was set to 0.5. Then, the p_{2i} 's were truncated so that they all lie in the interval $[0.5;0.8]$. The average of the p_{2i} 's was equal to 65%.

Mechanism 3: The population U was divided into four groups U_1, \dots, U_4 , according to the quartiles of the x_2 -values. Then, a probability p_{3i} was assigned to unit i such that $p_{3i} = 0.5$ if $i \in U_1$, $p_{3i} = 0.6$ if $i \in U_2$, $p_{3i} = 0.7$ if $i \in U_3$ and $p_{3i} = 0.8$ if $i \in U_4$. The average of the p_{3i} 's was equal to 65%.

Note that for mechanisms 1 and 2, the probability of e-filer/p-filer status depends on the variable of interest y_1 but is independent of y_2 and y_3 . For mechanism 3, the probability of e-filer/filer status depends on the variable x_2 only. Finally, in each population, $B = 500$ e-filers indicators variables $a_{1i}^{(b)}$, $a_{2i}^{(b)}$ and $a_{3i}^{(b)}$, $b = 1, \dots, B$, were generated independently from Bernoulli distributions with probability p_{1i} , p_{2i} and p_{3i} , respectively. Table 2 shows the population means for the population of e-filers and that of the p-filers for the 27 scenarios considered in this study.

Table 2: Mean values for the variables of interest for three response mechanisms

		\bar{Y}_E			\bar{Y}_P		
		y_1	y_2	y_3	y_1	y_2	y_3
Population	E-filers/p-filers						
	Mechanism						
1	1	24.0	188.3	10.0	15.4	151.3	10.0
	2	17.6	161.8	10.0	26.8	199.1	10.0
	3	22.0	202.5	10.0	19.0	124.9	10.0
	1	23.4	190.4	9.9	16.9	145.1	9.9

Population	2	18.7	158.2	9.9	25.3	202.9	9.9
2	3	22.2	201.4	9.9	19.1	124.7	9.9
Population	1	22.9	193.2	10.0	17.5	138.3	10.0
	2	19.1	155.1	10.0	24.4	206.5	10.0
	3	22.1	200.8	10.0	19.1	124.2	10.0

From each population and for each of the 500×3 e-filers indicators variables, a sample of size $n = 500$ according to three sampling procedures: (i) simple random sampling from the stratum of e-filers. Under this procedure, we computed the Hajek estimator (HAJ) given by (1.1), the direct calibrated estimator (DCAL) given by (2.4) with $\mathbf{x} = (1, x_1, x_2)'$ and the generalized calibrated estimator (GCAL) given by (2.16) with $\mathbf{x} = (1, x_1, x_2)'$ and $\mathbf{z} = (1, x_1, y_1)'$. (ii) Balanced sampling as described in section 3.2. We computed the corrected balanced estimator (CBAL) given by (3.6) with $\mathbf{x} = (1, x_1, x_2)'$. (iii) Balanced sampling as described in section 3.3. We computed the corrected after reweighting balanced estimator (CARBAL) given by (3.9) with $\mathbf{x} = (1, x_1, x_2)'$ and $\mathbf{z} = (1, x_1, x_2)'$. Note that the variable y_1 is not included in the vector \mathbf{z} of instrumental variables for CARBAL, since it is unknown at the design stage. Consequently, the CARBAL strategy considered in this simulation is identical to the CBAL strategy, and is not included in the results presented below. Also, to evaluate the performance of the methods when relevant explanatory variables are omitted, we conducted the two first procedures by using DCAL with $\mathbf{x} = (1, x_1)'$, GCAL with $\mathbf{x} = (1, x_1)'$ and $\mathbf{z} = (1, y_1)'$ and CBAL with $\mathbf{x} = (1, x_1)'$.

As a measure of bias of a point estimator $\hat{\theta}$ of parameter θ , we used the Monte Carlo percent relative bias (RB) given by

$$RB_{MC}(\hat{\theta}) = 100 * \frac{B^{-1} \sum_{b=1}^B (\hat{\theta}_{(b)} - \theta)}{\theta},$$

where $\hat{\theta}_{(b)}$ denotes the estimator $\hat{\theta}$ in the b^{th} sample. As a measure of stability of an estimator $\hat{\theta}$, we used the Monte Carlo percent relative root mean square error (RRMSE) given by

$$RRMSE_{MC}(\hat{\theta}) = 100 * \frac{\sqrt{B^{-1} \sum_{b=1}^B (\hat{\theta}_{(b)} - \theta)^2}}{\theta}.$$

Tables 3, 4 and 5 show the Monte Carlo percent RB and RRMSE (in brackets), corresponding to the variables y_1, y_2 and y_3 , respectively.

From Table 1 (corresponding to the variable y_1), we first note that the HAJ estimator shows an appreciable RB in all the scenarios. This result is not surprising since the HAJ estimator does not make use of any auxiliary information that is either related to y_1 nor to the probability of e-filers/p-filer status. As a result, the HAJ shows a large RRMSE. We now discuss the results obtained when the complete auxiliary information is used. For the DCAL estimator, we note that RB is negligible for mechanism 3 for the three populations. This can be explained by the fact that the variable x_2 is related to both the e-filer/p-filer status and the y_1 -variable and that x_2 was used as one of the calibration variable. For mechanisms 1 and 2, we note that the RB does not vanish but is significantly smaller than that of HAJ estimator. The residual bias can be explained by the fact that the e-filer/p-filer status depends on the variable y_1 . Also, we note that the bias decreases as the predictive power of the model

(quantified by the model R^2) increases. Turning to the GCAL estimator that uses the variable y_1 as an instrumental variable, we note that the bias was virtually eliminated for mechanisms 1 and 2. This example clearly illustrates the advantage of generalized calibration when the e-filer/p-filer status depends on the variable of interest being estimated. However, it is worth noting that for mechanism 3, the GCAL estimator is heavily biased for all the populations. For example, the absolute RB of the GCAL estimator for population 1 is approximately equal to 24%, which is almost five times the bias of the naïve HAJ estimator. This can be explained by the fact that, under mechanism 3, the e-filer/p-filer status depends on x_2 and that x_2 was not included in the \mathbf{z} -vector for the GCAL estimator. As a result, there remains a residual relationship between the e-filer/p-filer and \mathbf{x} after conditioning on \mathbf{z} . Comparing the DCAL and the CBAL estimators, we note that they lead to very similar results in terms of both RB and RRMSE, which is not surprising since both estimators make use of the same auxiliary information. When some of the auxiliary information is omitted, the gain of the DCAL and the CBAL estimators is more limited, as expected. For the GCAL estimator, we obtain similar results for mechanisms 1 and 2, but for mechanism 3, the GCAL estimator shows a small RB. This can be explained by the fact that, even though the e-filer/p-filer status depends on x_2 , the variable x_2 is not part of the \mathbf{x} -vector. As a result, there is no residual relationship between the e-filer/p-filer and \mathbf{x} after conditioning on \mathbf{z} .

Similar conclusions may be drawn for the variable y_2 (see Table 4). The HAJ estimator is heavily biased, and shows a large RRMSE. For the DCAL estimator, we note that the RB is larger than that obtained for the variable y_1 for the mechanisms 1 and 2. For the mechanism 3, the bias does not vanish (as it did for the variable y_1), which can be explained by the fact that the quadratic term was omitted from the vector of calibration variables. Despite the

omission of x_2 , the RB obtained with the DCAL estimators are significantly smaller than that obtained with HAJ. Once again, the RB decreases as the predictive power increases. Turning to the GCAL, we note the RB is negligible for mechanisms 1 and 2 when the variable y_1 is included in the instrumental variables. However, the RB is substantial for mechanism 3, for the same reasons explained above. These results clearly illustrate the need for a careful work for describing the response mechanism.

For the variable y_3 (which is unrelated to all the variables), all the estimators are virtually unbiased, as expected (see Table 5). In terms of RRMSE, the results are similar for the five estimators although we note a slight loss of efficiency for the DCAL and GCAL estimators.

5. SUMMARY AND DISCUSSION

In this paper, we studied the problem of sampling and estimation in the context of cut-off sampling. We showed that naïve point estimators could be severely biased if the units excluded from the sample are significantly different from the rest of the population. This situation is not uncommon in practice. In order to reduce the bias, we considered two well known techniques, namely balanced sampling and calibration. From a bias point of view, generalized calibration is particularly interesting because it allows for the use of all the auxiliary information (that related to the variable of interest and that explaining the e-filer/p-filer status) as well as the inclusion of the variables of interest in the vector of auxiliary variables. Moreover, unlike balanced sampling, generalized calibration is performed at the estimation stage so the auxiliary information available at this stage is typically richer than the one available at the sampling stage. In practice, it might be wise to select a sample balanced on the design variables and to use calibration to satisfy control totals corresponding to the calibration variables and/or design variables.

ACKNOWLEDGMENTS

The authors would like to thank an Associate Editor and an anonymous referee for their useful comments and suggestions that contributed to significantly improving the paper. Work of David Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

- Deville, J-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Recueil de la section des méthodes d'enquête*, Congrès de la Société Statistique du Canada, Sherbrooke, 103-110.
- Deville, J-C. (2002). La correction de la non-réponse par calage généralisé. *Actes des Journées de Méthodologie Statistique*, Insee.
- Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators In Survey Sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Deville, J-C. and Särndal, C-E., and Sautory, O. (1993). Generalized Raking procedure in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- Deville, J-C. and Tillé, Y. (2004). Efficient balanced sampling : the Cube method. *Biometrika*, 91, 893-912.
- Deville, J-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.
- Fecteau, S. and Jocelyn, W. (2005). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incorporées. *Actes du Colloque Francophone sur les Sondages*, Québec.
- Hajek, J. (1971). Comment on a paper of D. Basu, in Godambe, V.P. Sprott, D.A., Eds, *Foundations of Statistical Inference*, Toronto, Holt, Rhinehart and Winston, 236.
- Kott, P. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. *Survey Methodology*, 32, 133-142.
- Le Guennec, J. and Sautory, O. (2002). Application du calage généralisé à la correction de la non-réponse : une expérimentation. *Actes des Journées de Méthodologie Statistique*, Insee.
- Sautory, O. (2003). Calmar 2: a new version of the Calmar calibration adjustment program. *Proceedings of the Statistics Canada Symposium, Ottawa, Canada*.
- Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing*. Chapman & Hall, London.

Table 3: Monte Carlo percent RB and Monte Carlo RRMSE for seven estimators of the population total Y_1

	Mechanism	HAI	DCAL $\mathbf{x} = (1, x_1, x_2)'$	GCAL $\mathbf{x} = (1, x_1, x_2)'$ $\mathbf{z} = (1, x_1, y_1)'$	CBAL $\mathbf{x} = (1, x_1, x_2)'$	DCAL $\mathbf{x} = (1, x_1)'$	GCAL $\mathbf{x} = (1, x_1)'$ $\mathbf{z} = (1, y_1)'$	CBAL $\mathbf{x} = (1, x_1)'$
Population 1 ($R^2 = 0.3$)	1	13.8 (14.3)	9.8 (10.3)	0.5 (7.5)	9.9 (10.4)	11.8 (12.3)	0.5 (9.5)	11.6 (12.1)
	2	-15.0 (15.4)	-10.6 (11.1)	-0.4 (8.0)	-10.7 (11.1)	-12.7 (13.1)	0.2 (9.1)	-12.7 (13.2)
	3	4.7 (6.1)	-0.1 (3.3)	-24.1 (25.6)	-0.1 (3.0)	4.7 (5.9)	4.6 (9.9)	5.1 (6.1)
Population 2 ($R^2 = 0.5$)	1	10.7 (11.2)	5.7 (6.0)	0.4 (3.7)	5.6 (6.0)	8.2 (8.6)	0.5 (5.1)	8.3 (8.7)
	2	-11.5 (11.8)	-6.0 (6.4)	-0.5 (4.0)	-6.0 (6.4)	-8.7 (9.0)	-0.6 (5.4)	-8.7 (9.0)
	3	5.3 (6.0)	0.0 (2.1)	-10.8 (11.4)	0.2 (2.1)	5.1 (5.7)	4.8 (7.2)	5.1 (5.7)
Population 3 ($R^2 = 0.7$)	1	8.9 (9.2)	2.9 (3.2)	0.3 (1.9)	3.0 (3.3)	6.0 (6.3)	0.2 (3.1)	6.1 (6.4)
	2	-9.0 (9.3)	-2.9 (3.2)	-0.2 (1.9)	-2.8 (3.1)	-5.9 (6.2)	-0.1 (3.6)	-5.7 (6.0)
	3	5.1 (5.6)	0.1 (1.3)	-4.3 (4.7)	0.0 (1.3)	5.0 (5.4)	5.2 (6.3)	5.0 (5.4)

Table 4: Monte Carlo percent RB and Monte Carlo RRMSE for seven estimators of the population total Y_2

	Mechanism	HAJ	DCAL $\mathbf{x} = (1, x_1, x_2)'$	GCAL $\mathbf{x} = (1, x_1, x_2)'$ $\mathbf{z} = (1, x_1, y_1)'$	CBAL $\mathbf{x} = (1, x_1, x_2)'$	DCAL $\mathbf{x} = (1, x_1)'$	GCAL $\mathbf{x} = (1, x_1)'$ $\mathbf{z} = (1, y_1)'$	CBAL $\mathbf{x} = (1, x_1)'$
Population 1 ($R^2 = 0.3$)	1	36.5 (38.1)	28.6 (30.1)	0.8 (22.4)	28.8 (30.3)	36.5 (38.1)	1.2 (29.9)	35.7 (37.4)
	2	-39.5 (40.9)	-31.5 (33.0)	-0.9 (23.5)	-31.2 (32.6)	-39.1 (40.5)	1.5 (29.5)	-39.2 (40.6)
	3	14.4 (18.5)	-3.8 (10.6)	-75.2 (76.6)	-3.8 (10.1)	14.4 (18.5)	14.4 (31.2)	15.8 (19.3)
Population 2 ($R^2 = 0.5$)	1	25.9 (27.5)	15.8 (17.1)	0.5 (11.4)	15.7 (16.9)	25.9 (27.5)	0.8 (16.6)	26.3 (27.8)
	2	-27.7 (29.0)	-17.0 (18.4)	-0.6 (11.7)	-16.9 (18.2)	-27.0 (28.3)	-0.6 (18.4)	-27.2 (28.3)
	3	16.0 (18.4)	-3.6 (7.3)	-35.3 (37.2)	-2.9 (7.0)	16.0 (18.4)	15.2 (23.8)	16.1 (18.5)
Population 3 ($R^2 = 0.7$)	1	19.3 (20.6)	7.5 (8.6)	-0.2 (5.9)	7.8 (8.9)	19.5 (20.8)	-0.2 (10.5)	19.8 (21.1)
	2	-19.3 (20.5)	-7.4 (8.9)	0.4 (5.9)	-7.0 (8.4)	-18.7 (19.9)	1.4 (13.5)	-17.9 (19.1)
	3	15.6 (17.4)	-3.4 (5.5)	-16.4 (17.4)	-3.6 (5.6)	15.6 (17.4)	16.3 (21.0)	15.7 (17.5)

Table 5: Monte Carlo percent RB and Monte Carlo RRMSE for seven estimators of the population total Y_3

	Mechanism	HAI	DCAL $\mathbf{x} = (1, x_1, x_2)'$	GCAL $\mathbf{x} = (1, x_1, x_2)'$ $\mathbf{z} = (1, x_1, y_1)'$	CBAL $\mathbf{x} = (1, x_1, x_2)'$	DCAL $\mathbf{x} = (1, x_1)'$	GCAL $\mathbf{x} = (1, x_1)'$ $\mathbf{z} = (1, y_1)'$	CBAL $\mathbf{x} = (1, x_1)'$
Population 1 ($R^2 = 0.3$)	1	0.2 (2.8)	0.1 (2.8)	0.0 (2.9)	0.2 (3.1)	0.2 (2.8)	0.0 (2.9)	0.2 (3.2)
	2	-0.2 (2.9)	-0.2 (2.9)	-0.1 (3.0)	-0.2 (3.1)	-0.3 (2.9)	-0.1 (3.0)	-0.2 (3.0)
	3	0.1 (3.1)	0.0 (3.1)	-0.1 (3.3)	-0.2 (3.2)	0.1 (3.1)	0.1 (3.1)	0.1 (3.0)
Population 2 ($R^2 = 0.5$)	1	-0.1 (3.1)	-0.1 (3.1)	0.0 (3.2)	-0.3 (3.1)	-0.1 (3.1)	0.0 (3.1)	-0.1 (3.0)
	2	0.2 (3.3)	0.2 (3.3)	0.1 (3.3)	0.2 (3.1)	0.2 (3.3)	0.1 (3.3)	0.0 (3.0)
	3	0.1 (3.1)	0.2 (3.2)	0.4 (3.3)	0.1 (3.0)	0.1 (3.1)	0.1 (3.1)	0.0 (3.0)
Population 3 ($R^2 = 0.7$)	1	0.1 (3.1)	0.0 (3.1)	0.1 (3.1)	0.0 (3.1)	0.1 (3.1)	0.1 (3.2)	-0.2 (3.0)
	2	0.2 (3.1)	0.2 (3.1)	0.2 (3.1)	0.4 (3.0)	0.2 (3.1)	0.2 (3.1)	0.2 (3.0)
	3	0.1 (3.2)	0.0 (3.3)	0.1 (3.4)	-0.1 (3.2)	0.1 (3.3)	0.1 (3.3)	0.0 (3.1)