INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2008-30

A Note on Sampling and Estimation in the Presence of Cut-Off Sampling

D. HAZIZA[1] **– G. CHAUVET**[2]
J-C. DEVILLE[3]

[1] Université de Montréal, Département de Mathématiques et de Statistique, Montréal, Québec H3C 3J7, Canada. david.haziza@umontreal.ca

[2] CREST-ENSAI, Laboratoire de Statistique d'Enquête, Campus de Ker Lann, 35170 Bruz, France. chauvet@ensai.fr

[3] CREST-ENSAI, Laboratoire de Statistique d'Enquête, Campus de Ker Lann, 35170 Bruz, France. deville@ensai.fr

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

David Haziza, Guillaume Chauvet and Jean-Claude Deville[1]

12 décembre 2008

Résumé

L'échantillonnage de type cut-off consiste à exclure délibérément de la sélection une partie des unités, par exemple si la contribution au total des unités exclues est faible et si l'inclusion de ces unités dans l'échantillon implique des coûts importants. Si les caractéristiques des unités exclues diffèrent de celles des unités dans l'ensemble de la population étudiée, l'utilisation d'estimateurs naïfs peut donner lieu à des estimations sévèrement biaisées. Dans cet article, nous discutons de l'utilisation d'information auxiliaire afin de réduire le biais de non-réponse par des techniques de calage ou d'échantillonnage équilibré. Nous montrons que l'utilisation d'information auxiliaire liée à la variable d'intérêt, et d'information auxiliaire liée à la probabilité de réponse, permet de réduire fortement le biais d'estimation. Une courte étude par simulations est également proposée.

Mots-clés : Biais sous le plan ; Biais sous le modèle ; Calage ; Echantillonnage de type cut-off ; Echantillonnage équilibré ; Information Auxiliaire.

Abstract

Cut-off sampling consists of deliberately excluding a set of units from possible samples selection, for example if the contribution of the excluded units to the total is small and if the inclusion of these units in the sample selection involves high costs. If the characteristics of the excluded units differ from that of the population under study, the use of naïve estimators may result in strongly biased estimations. In this paper, we discuss the use of auxiliary information to reduce the non-response bias by means of calibration or balanced sampling techniques. It is demonstrated that the use of both the available auxiliary information related to the variable of interest and of the available auxiliary information related to the probability of response enables to strongly reduce the estimation 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) et 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 sample selection, is frequently used in business surveys. The small businesses are often grouped into take-none strata and are thus excluded from possible sample selection. The contribution to the overall total of the excluded units is typically small. 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 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 $U = U_E \bigcup U_P$ and $N = N_E + N_P$. Fecteau and Jocelyn (2005) mention that the populations of e-filers and p-filers have different characteristics. Indeed, the businesses using a paper format are typically larger in size than the one 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 two populations differ with respect to several characteristics of interest, the use of naïve estimators may potentially lead to misleading results.

Auxiliary information plays an important role in surveys because it allows the survey statistician to use more efficient sampling and estimation procedures and can be used to reduce nonsampling errors. We distinguish between two sets of auxiliary variables: the first set is the set of design variables, we assume to be available for all the units in the population at the design stage. The design variables are typically used to stratify the population. Also, we assume that, at the estimation stage, a set of auxiliary variables (often called calibration variables), is available for all the sampled units and that the population total for each variable is known. Note that the two sets of auxiliary variables are not necessarily disjoint so a given auxiliary variable can be used at both stages (design and estimation).

In order to reduce the bias due to the exclusion of the p-filers from the sample, we consider two well known techniques that both use some amount of auxiliary information: (i) balanced sampling and (ii) calibration. One advantage of calibration over balanced sampling is that calibration can use both types of auxiliary variables (design and calibration variables), whereas balanced sampling requires the auxiliary variables to be available for all the units in the population. 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. 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.

In this paper, 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 $p_E(.)$ 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. Note that $\pi_i = 0$ for all $i \in U_P$. As a result, it is well known that a design-unbiased estimator of Y does not exist.

A basic estimator of Y is the so-called Hajek estimator given by

$$\hat{Y}_{HA} = N\bar{y}_E, \tag{1.1}$$

where $\bar{y}_E = \frac{\hat{Y}_E}{\hat{N}_E}$ with $\left(\hat{Y}_E, \hat{N}_E\right) = \sum_{i \in s_E} d_i (y_i, 1)$. The design-bias of $\hat{Y}_{HA}$ in (1.1) can be approximated by

$$\mathrm{B}_s\left(\hat{Y}_{HA}\right) \equiv E_s\left(\hat{Y}_{HA}\right) - Y \approx N_P\left(\bar{Y}_E - \bar{Y}_P\right), \tag{1.2}$$

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, and $E_s(.)$ denotes the expectation with respect to sampling. The bias in (1.2) is equal to 0 if $N_P = 0$, which occurs when $U_P = \varnothing$ or when $\bar{Y}_E = \bar{Y}_P$. These two conditions are not satisfied in practice since the population of p-filers represent 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 considerably biased when $\bar{Y}_E$ is significantly different than $\bar{Y}_P$. Note that $\bar{Y}_E = \bar{Y}_P$ when the variable of interest *y* is not related to the variable to the efiler/pfiler indicator variable. In general however, alternative strategies are 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 that use the auxiliary information available at the estimation stage. We study their properties in terms of bias. Balanced sampling as a mean to reduce the bias of the Hajek estimator is presented. We consider two versions of balanced sampling and discuss the properties of the resulting estimators. In Section 4, we conduct a limited simulation study to investigate on 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 the businesses by type of format

Business	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. We consider three classes of calibration estimators, which are presented in sections 2.1-2.3.

2.1 Direct calibration

Suppose that a vector of q auxiliary variables $\mathbf{x} = (x_1,...,x_q)'$ 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 $\mathbf{x}$ may be described according to the following model

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

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 q-vector of unknown parameters, σ^2 is an unknown parameter, c_i is a known constant and $E_m(.)$ and $V_m(.)$ denote the expectation and the variance with respect to the model (2.1). We assume that $c_i=\boldsymbol{\alpha}'\mathbf{x}_i,$ where $\boldsymbol{\alpha}$ is a q-vector of specified constants.

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

$$\sum_{i\in s_E} w_i\mathbf{x}_i=\mathbf{X} \tag{2.2}$$

is satisfied, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers and $F(.)$ is the so-called calibration function. Several calibration functions $F(.)$ can be used ; see Deville (2002) and Le Guennec and Sautory (2002). Two such options for $F(.)$ 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. 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 $F(.)$. The resulting calibration estimator of *Y* is given by

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

A special case of (2.3) is obtained under the linear function $F(u)=1+u,$ which leads to the generalized regression estimator

$$\hat{Y}_G=\sum_{i\in s_E} w_iy_i, \tag{2.4}$$

where

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

with $\hat{\mathbf{T}}_E = \sum_{i \in s_E} d_i c_i^{-1} \mathbf{x}_i \mathbf{x}_i'$ and $\hat{\mathbf{t}}_E = \sum_{i \in s_E} d_i c_i^{-1} \mathbf{x}_i y_i$. 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 design-bias of $\hat{Y}_G$ in (2.4) is given by

$$B_s\left(\hat{Y}_G\right) \equiv E_s\left(\hat{Y}_G\right) - Y \approx -\sum_{i \in U_P} \left(y_i - \mathbf{x}_i' \mathbf{B}_E\right), \tag{2.6}$$

where $\mathbf{B}_E = \left[\sum_{i \in U_E} c_i^{-1} \mathbf{x}_i \mathbf{x}_i'\right]^{-1} \sum_{i \in U_E} c_i^{-1} \mathbf{x}_i y_i$ denotes the census coefficient of regression corresponding to the population of e-filers. The asymptotic design-bias of $\hat{Y}_G$ in (2.6) is small if the residuals $E_i = \left(y_i - \mathbf{x}_i' B_E\right)$ corresponding to the p-filers are small, which in turns suggests that the model (2.1) holds for the p-filers. On the other hand, $\hat{Y}_G$ is asymptotically *sm*-unbiased under the model (2.1). That is, $E_s E_m \left(\hat{Y}_G - Y\right) \approx 0$. However, since no *y*-value is 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

$$B_{sm}\left(\hat{Y}_G\right) \approx \mathbf{X}_P' \left(\boldsymbol{\beta}_E - \boldsymbol{\beta}_P\right), \tag{2.7}$$

where $\mathbf{X}_P = \sum_{i \in U_P} \mathbf{x}_i$. 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 is available for the p-filers. This example illustrates the problematic 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 $\hat{Y}_{HA}$.

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. Let a_i be a binary variable such that $a_i = 1$ if unit *i* is an e-filer and $a_i = 0,$ otherwise. We assume that a_i is available for all the units in the population. Also, let $p_i = P(a_i = 1)$ be the probability that a unit is an e-filer.

In addition to the vector of auxiliary variables $\mathbf{x}_i$ related to the variable of interest y_i by the model (2.1), we assume that there exists a *l*-vector of auxiliary variables $\mathbf{z}_i$ related to the probability p_i. 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 model:

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

where $f(.)$ is a given function, and $\boldsymbol{\gamma}$ is a vector of unknown parameters. A special case of (2.8) is the logistic regression model given by

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

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 $\boldsymbol{\gamma}$).

A second set of estimators can be obtained by finding a new set of weights w_i^* so that the calibration equation

$$\sum_{i\in s_E} d_i^* \mathbf{x}_i F(\boldsymbol{\lambda}'\mathbf{x}_i) = \mathbf{X}, \tag{2.10}$$

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

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

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

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

where $w_i^* = d_i^*\left(1 + c_i^{-1}\left(\mathbf{X} - \hat{\mathbf{X}}_E^*\right)' \hat{\mathbf{T}}_E^{*-1}\mathbf{x}_i\right)$ with $\hat{\mathbf{T}}_E^* = \sum_{i\in s_E} d_i^* c_i^{-1}\mathbf{x}_i\mathbf{x}_i'$ and $\left(\hat{\mathbf{X}}_E^*, \hat{Y}_E^*\right) = \sum_{i\in s_E} d_i^*\left(\mathbf{x}_i, y_i\right)$.

If the assumed model (2.8) is valid (i.e., $\hat{p}_i \approx p_i$), then the alternative regression estimator $\hat{Y}_G^*$ (2.12) is asymptotically $s\xi$-unbiased. That is, $E_s E_\xi\left(\hat{Y}_G^*\right) \approx Y$. On the other hand, if the model (2.1) holds, then $\hat{Y}_G^*$ is asymptotically *sm*-unbiased. That is, $E_s E_m\left(\hat{Y}_G^* - Y\right) \approx 0$. Hence, $\hat{Y}_G^*$ is doubly robust in the sense that it is valid if one model or the other 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 all the appropriate auxiliary information ($\mathbf{x}$ and $\mathbf{z}$) available, unlike the estimator $\hat{Y}_G$ given by (2.4) that only uses the auxiliary information $\mathbf{x}$. As a result, it is expected that the use of $\hat{Y}_G^*$ will achieve an effective bias

reduction if either the model *m* or the model 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 $s\xi$-unbiased for any variable of interest *y*, as long as the probability $\hat{p}_i$ is 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.

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 using the so-called generalized calibration (Deville, 1998; 2002 and Kott, 2006). It assumes that the vector $\mathbf{x}$ 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 it can be used if the values of the $\mathbf{z}$ variables are only known for the e-filers. 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. This is illustrated in the simulation study presented in section 4. We seek for a new set of weights $\tilde{w}_i$ so that the calibration equation

$$\sum_{i\in s_E} d_i \mathbf{x}_i F(c_i^{-1}\boldsymbol{\lambda}'\mathbf{z}_i) = \mathbf{X} \tag{2.13}$$

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

$$\hat{\hat{Y}}_{CAL} = \sum_{i\in s_E} d_i F(c_i^{-1}\boldsymbol{\lambda}'\mathbf{z}_i) y_i. \tag{2.14}$$

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

$$\hat{\hat{Y}}_G = \sum_{i\in s_E} \tilde{w}_i y_i, \tag{2.15}$$

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

$$\hat{\tilde{Y}}_G = \hat{Y}_E + \left(\mathbf{X} - \hat{\mathbf{X}}_E\right)' \hat{\tilde{\mathbf{B}}}_E, \tag{2.16}$$

where $\hat{\tilde{\mathbf{B}}}_E = \hat{\tilde{\mathbf{T}}}_E^{-1} \hat{\tilde{\mathbf{t}}}_E$ with $\hat{\tilde{\mathbf{t}}}_E = \sum_{i \in s_E} d_i c_i^{-1} \mathbf{z}_i y_i$. Thus, the estimated regression coefficient $\hat{\tilde{\mathbf{B}}}_E$ can be viewed as the estimated regression coefficient obtained by fitting an instrumental regression analysis with the vector $\mathbf{z}$ as the instruments. If the vector $\mathbf{z}$ explains well the fact of being an e-filer, then 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 $\tilde{\mathbf{T}}_E = \sum_{i \in U_E} c_i^{-1} \mathbf{x}_i \mathbf{z}_i'$ may be close to singularity, which may result in a highly variable calibrated estimator $\hat{\tilde{Y}}_{CAL}$. Note that when $\mathbf{z}_i = \mathbf{x}_i$, the estimator (2.14) reduces to the direct calibration estimator given by (2.3).

3. BALANCED SAMPLING

A balanced sampling design ensures that Horvitz-Thompson estimators of the auxiliary variables, called balancing variables, exactly match the known totals. The Horvitz-Thompson estimator is still design-unbiased, while its variance is only given by a residual variable associated to a regression of the variable of interest on balancing variables, and is thus strongly reduced if these balancing variables are well correlated with the variable of interest. 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.

3.1 The traditional balanced sampling

In this paragraph, a brief introduction to the Cube method (see Deville and Tillé, 2004) is given. We shall first assume that the stratum U_p is empty, that is, that the sample is directly selected in population U with non zero inclusion probability π_i for unit i, and we simply note $s_E = s$. Assume that the vectors of values $\mathbf{x}_i = \left(x_{i1}, \cdots, x_{iq}\right)'$ taken by q auxiliary variables are known for all the units of the population. Recall that $\hat{\mathbf{X}} = \sum_{i \in s} \mathbf{x}_i / \pi_i$ is an unbiased estimator of $\mathbf{X} = \sum_{i \in U} \mathbf{x}_i$; that is,

$$E_s\left(\hat{\mathbf{X}}\right) = \mathbf{X}. \tag{3.1}$$

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

$$\hat{\mathbf{X}} = \mathbf{X} \tag{3.2}$$

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 impose a fixed sample size. In the special case of $\mathbf{x}_i = 1$, $\hat{\mathbf{X}}$ and $\mathbf{X}$ represent the estimated and exact population size respectively and (3.2) is equivalent to give an exact estimation of the population size.

The Cube method proposed by Deville and Tillé (2004) provides a general algorithm for selecting balanced samples with predetermined inclusion probabilities. As an exact balanced sampling design generally does not exist, that is, there may exist no sample such that equation (3.2) holds, the objective is generally to select a sample such that the balancing equations

(3.2) holds approximately. 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 equation are exactly respected. This step stops when the balancing conditions may no more be exactly respected. The last step consists in ending the sampling so that the inclusion probabilities remain exactly respected and the balancing conditions remain approximately respected. In the context of balanced sampling, note that the balancing variables $\mathbf{x}_i$ for all the population units prior to sampling, whereas only the $\mathbf{X}$ totals and the $\mathbf{x}_i$ values for the sampled units are needed in the context of calibration.

3.2 The corrected balanced sampling

We now turn back to the general setting of non empty stratum U_P. We assume that a vector of q auxiliary variables $\mathbf{x} = (x_1, \ldots, x_q)'$ is available for all the units in the population U (e-filers and p-filers), and 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). Assume that the sample s_E is selected in U_E by balanced sampling with inclusion probability π_i for unit $i \in U_E$ and balancing variables $\mathbf{x}$, so that the equation

$$\sum_{i \in s_E} \frac{\mathbf{x}_i}{\pi_i} = \sum_{i \in U_E} \mathbf{x}_i \qquad (3.3)$$

holds. The Hajek estimator $\hat{Y}_{HA} = N\overline{y}_E$ given by (1.1) may then be used for estimating the total Y, but this estimator remains design-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

$$\frac{1}{n}\sum_{i\in s_E}\mathbf{x}_i = \overline{\mathbf{X}} \tag{3.4}$$

are satisfied, where $\overline{\mathbf{X}} = \frac{\mathbf{X}}{N}$ denotes the overall population mean of the $\mathbf{x}$- vector. Condition (3.4) may be obtained in the following way. 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

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

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) give

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

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

$$\hat{Y}_{sm} = N\tilde{\tilde{y}}_E \tag{3.6}$$

where $\tilde{\tilde{y}}_E = \frac{1}{n}\sum_{i\in s_E} y_i$ is the sample mean of y. The estimator $\hat{Y}_{sm}$ in (3.6) is called the corrected balanced estimator, and is asymptotically m-unbiased under the model (2.1). Note that this estimator makes no use of the design weights.

3.3 Balanced sampling after reweighting

In this section, we propose an estimator obtained by a modified balanced sampling design, and that 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

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

for some function $H(.)$. We seek for 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

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

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

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

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 $s\xi$-unbiased if the model (2.9) is valid, and also asymptotically sm-unbiased if the model (2.1) is valid. Note that if s_E is selected in U_E with probabilities π_i proportional to $1/p_i$, the design weights are not needed.

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 a finite population of size $N = 10\,000$ containing 4 variables: two variables of interest y_1 and y_2 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 parameters 2 and 5. Given the x_1-values and the x_2-values, the y_1-values were generated according to the linear regression model

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

The η_i's were generated according to a normal distribution with mean 0 and variance σ^2. The model parameters β_0, β_1 and β_2 were all set to 1 and the variance σ^2 was chosen to give a model R^2 (coefficient of determination) approximately equal to 0.7. Finally, the y_2-values were generated independently from a Gamma distribution with parameters 2 and 5.

The population was partitioned into the strata of e-filers and the strata of p-filers as follows: first, probabilities were assigned to each population unit according to two mechanisms:

Mechanism 1: The population U was divided into four groups $U_1,\cdots,U_4$, according to the quartiles of the y_1-values. Then, we attach a probability p_{1i} 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 $i\in U_4$. 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

$$p_{2i}=\exp(-\gamma_0\, y_{1i}/\bar{Y}_1), \tag{4.2}$$

where $\bar{Y}_1$ denotes the population mean of variable y_1. The parameter γ_0 was set to 0.5. Then, the p_{2i}'s were truncated to be included between 0.5 and 0.8. The average of the p_{2i}'s was equal to 63.5%.

Note that for both mechanisms 1 and 2, the probability of being an e-filer depends on the variable of interest y_1 but is independent of y_2. Finally, the e-filers indicators variables a_{1i} and a_{2i} were generated independently from Bernoulli distributions with probability p_{1i} and p_{2i}, respectively leading to two different partitions of the original populations. The objective is to estimate the population totals of the variables of interest y_1 and y_2, denoted by Y_1 and Y_2, respectively.

We selected $B=1000$ samples of size $n=500$ according to three 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 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 $\mathbf{z} = (1, x_1, x_2)'$.

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

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

where $\hat{\theta}_{(b)}$ gives the value of the estimator for the b^{th} sample. As a measure of variance of an estimator $\hat{\theta}$, we used the monte carlo percent relative root mean square error (RRMSE) given by

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

The monte carlo percent RB and RRMSE are shown in Tables 2 and 3, respectively. We first discuss the results corresponding to the variable y_1. For both mechanisms 1 and 2, the HAJ estimator shows an appreciable RB (approximately 8.8% for mechanism 1 and – 9% for mechanism 2). This result is not surprising since the HAJ estimator does not make use of any auxiliary information that is either related to y_1 or to p_1 (for mechanism 1) and p_2 (for mechanism 2). As a result, the HAJ shows a large RRMSE. For the DCAL and CBAL estimators that use the auxiliary information that is related to y_1, we note that the RB is smaller than the one obtained for the HAJ estimator (approximately 2.8% for mechanism 1 and – 2.9% for mechanism 2 for the DCAL estimator). The RB values are very similar for

both estimators. Turning to the GCAL and CARBAL estimators that use auxiliary information as well as y_1 as an instrumental variable, we note that the bias was virtually eliminated (approximately 0.2% for mechanism 1 and – 0.2% for mechanism 2). This result shows the importance of using all the available auxiliary information as well as the variables of interest. In terms of RRMSE, the GCAL and CARBAL estimators show the lowest RRMSE. This can be explained that in both cases, the bias was virtually eliminated.

For the variable y_2 (that is not related to any variable), all five estimators are virtually unbiased, as expected. In terms of RRMSE, the results are similar for the five estimators although we note a slight loss of efficiency for the DCAL, GCAL and CARBAL estimators.

Table 2: Monte carlo percent RB for five estimators

	HAJ	DCAL	GCAL	CBAL	CARBAL
Variable of interest	*Mechanism 1*				
y_1	8.62	2.80	0.27	2.82	2.85
y_2	-0.34	-0.31	-0.22	-0.32	-0.17
	Mechanism 2				
y_1	-8.97	-2.78	-0.12	-2.87	-2.93
y_2	0.25	0.23	0.15	0.07	0.23

Table 3: Monte carlo percent RRMSE for five estimators

	HAJ	DCAL	GCAL	CBAL	CARBAL
Variable of interest	Response probability p_1				
y_1	8.94	3.10	1.86	3.12	3.16
y_2	3.21	3.24	3.25	3.09	3.06
	Response probability p_2				
y_1	9.28	3.08	1.86	3.16	3.22
y_2	3.04	3.10	3.11	3.01	3.12

5. SUMMARY AND DISCUSSION

In this paper, we studied the problem of sampling and estimation in the context of cut-of 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. Form a bias point of view, generalized calibration is particularly interesting because it allows for the use of all the auxiliary information (the one related to the variable of interest and the one explaining the probability of being excluded from the sample) as well as the variables of interest themselves that can be included in the auxiliary vector as instrumental 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.

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., 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.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and undercoverage. Survey Methodology, 32, pp. 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.