Statistica Sinica **27** (2017), 313-334
doi:http://dx.doi.org/10.5705/ss.2013.244

EXAMINING SOME ASPECTS OF BALANCED SAMPLING IN SURVEYS

Guillaume Chauvet, David Haziza and Éric Lesage

ENSAI/IRMAR, Université de Montréal and INSEE

Abstract: Balanced sampling has received some attention in recent years. There exists a number of procedures leading to a balanced or approximately balanced sample. In this paper, we examine the design-based properties of several estimation procedures under balanced sampling. The results of an extensive simulation study that compares different estimators are presented.

Key words and phrases: Cube algorithm, design-based inference, greg estimator, inclusion probability, Monte Carlo approximation, rejective sampling.

1. Introduction

Balanced sampling has received some attention in recent years, e.g., Deville and Tillé (2004), Chauvet and Tillé (2006), Fuller (2009) and Legg and Yu (2010). Consider a finite population P of size N. We are interested in estimating the population total $t_y = \sum_{i \in P} y_i$, where y denotes a characteristic of interest. Prior to sampling, we assume that a J-vector of auxiliary variables, $\mathbf{z}$, is available for all $i \in P$. The $\mathbf{z}$-variables are often referred to as the design variables.

A sample $s \subset P$ is said to be $\psi\mathbf{z}$-balanced if

$$\widehat{\mathbf{t}}_{\mathbf{z}}^{\psi} \equiv \sum_{i \in s} \psi_i^{-1} \mathbf{z}_i = \sum_{i \in P} \mathbf{z}_i \equiv \mathbf{t}_{\mathbf{z}}, \tag{1.1}$$

where $0 < \psi_i < 1$ for all $i \in P$. A design satisfying (1.1) for all s is called a $\psi\mathbf{z}$-balanced sampling design. There exists a number of procedures leading to a balanced or approximately balanced sample, including the Cube method (Deville and Tillé (2004)) and rejective sampling (Hájek (1981); Fuller (2009)).

Let π_i denote the inclusion probability attached to unit i with respect to the sampling design used to select the random sample S. The Cube method consists of assigning an inclusion probability π_i to every population unit prior to sampling and selecting a sample so that

$$\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi} \equiv \sum_{i \in s} \pi_i^{-1} \mathbf{z}_i = \mathbf{t}_{\mathbf{z}}. \tag{1.2}$$

The $\mathbf{z}$-variables in (1.1) and (1.2) are referred to as the balancing variables. The Cube method ensures that the π_i's are exactly satisfied. However, there may not exist a sample s exactly $\pi\mathbf{z}$-balanced, and the balancing constraints may be only approximately satisfied. This is referred to as the rounding problem.

In contrast, rejective sampling consists of selecting repeated samples according to a basic sampling procedure $p_b(\cdot)$ with basic inclusion probabilities $p_i = pr(i \in S_b)$ until

$$(\widehat{\mathbf{t}}^p_{\mathbf{z}} - \mathbf{t}_{\mathbf{z}})^\top V_b(\widehat{\mathbf{t}}^p_{\mathbf{z}})^{-1}(\widehat{\mathbf{t}}^p_{\mathbf{z}} - \mathbf{t}_{\mathbf{z}}) \leq \gamma^2, \tag{1.3}$$

where $\widehat{\mathbf{t}}^p_{\mathbf{z}} = \sum_{i\in S_b} p_i^{-1}\mathbf{z}_i$ with S_b denoting a random sample selected according to the basic procedure and $V_b(.)$ denotes the variance operator with respect to the basic procedure. The coefficient $\gamma > 0$ in (1.3) is a balancing tolerance specified by the survey statistician. The resulting rejective sampling procedure $p(\cdot)$ and the associated random sample S are not to be confused with $p_b(\cdot)$ and S_b, since in particular

$$\pi_i \equiv pr(i \in S) = pr\left(i \in S_b \,\middle|\, (\widehat{\mathbf{t}}^p_{\mathbf{z}} - \mathbf{t}_{\mathbf{z}})^\top V_b(\widehat{\mathbf{t}}^p_{\mathbf{z}})^{-1}(\widehat{\mathbf{t}}^p_{\mathbf{z}} - \mathbf{t}_{\mathbf{z}}) \leq \gamma^2\right) \neq p_i.$$

There is an important distinction between the Cube method and the rejective method. In the first, the inclusion probabilities π_i are exactly satisfied but one has no control on the (possible) discrepancy between the estimates $\widehat{\mathbf{t}}^\pi_{\mathbf{z}}$ and the true population totals $\mathbf{t}_{\mathbf{z}}$. In the second, the discrepancy between $\widehat{\mathbf{t}}^p_{\mathbf{z}}$ and $\mathbf{t}_{\mathbf{z}}$ is perfectly controlled through the balancing tolerance γ but the inclusion probabilities π_i are usually unknown.

In this paper, we examine the design-based properties of some estimation procedures with respect to the Cube method and rejective sampling. We adopt the following notation. Let $\mathbf{I} = (I_1, \ldots, I_N)^\top$ be the vector of sample selection indicators, and let $\mathbf{y} = (y_1, \ldots, y_N)^\top$ be the vector of the population y-values. In the design-based approach, the properties of estimators are evaluated with respect to the sampling design: the vectors $\mathbf{y}$ and $\mathbf{z}$ are held fixed and the only remaining source of randomness is the vector of sample indicators $\mathbf{I}$.

As an estimator of t_y, we consider linear estimators of the form

$$\widehat{t}^w_y = \sum_{i\in S} w_i y_i, \tag{1.4}$$

where w_i is a weight attached to unit i. The weights $w_i = \pi_i^{-1}$ lead to the customary Horvitz-Thompson estimator, $\widehat{t}^\pi_y$; it is design-unbiased and design-consistent for t_y, regardless of the y-variable being estimated, provided that the inclusion probabilities π_i's are known without error. If the π_i's are unknown (as for rejective sampling), we may settle for some approximation $\widehat{\pi}_i$, say. If $\widehat{\pi}_i$ provides a good approximation of π_i, we expect (1.4) based on $w_i = \widehat{\pi}_i^{-1}$ to exhibit a small bias.

Now, suppose that a q-vector of calibration variables $\mathbf{u} = (u_1, \ldots, u_q)^\top$ is available at the estimation stage for all the sample units, and that the vector of population totals $t_{\mathbf{u}} = \sum_{i \in U} \mathbf{u}_i$ is known. An alternative set of weights is

$$w_i = \pi_i^{-1} \Big\{ 1 + \big(t_{\mathbf{u}} - \widehat{t}_{\mathbf{u}}^{\pi} \big)^\top \Big(\sum_{i \in s} \pi_i^{-1} \mathbf{u}_i \mathbf{u}_i^\top \Big)^{-1} \mathbf{u}_i \Big\}, \tag{1.5}$$

where $\widehat{t}_{\mathbf{u}}^{\pi} = \sum_{i \in s} \pi_i^{-1} \mathbf{u}_i$. The choice (1.5) leads to the Generalized Regression (GREG) estimator noted as $\widehat{t}_{reg}^{\pi}$. The GREG estimator can be constructed with the assistance of the model

$$\begin{aligned} y_i &= \mathbf{u}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad E_m(\epsilon_i | \mathbf{u}) = 0, \\ V_m(\epsilon_i | \mathbf{u}) &= \sigma^2, \quad E_m(\epsilon_i \epsilon_j | \mathbf{u}) = 0, i \neq j, \end{aligned} \tag{1.6}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters and the subscript m denotes Model (1.6). Once again, provided that the inclusion probabilities π_i's are known without error, the GREG estimator is asymptotically design-unbiased and design-consistent for t_y regardless of the y-variable being estimated. This holds true even if Model (1.6) is misspecified; e.g., Särndal, Swensson and Wretman (1992). Now, suppose that the π_i's are unknown and are replaced by some approximation $\widehat{\pi}_i$ in (1.5). If $\widehat{\pi}_i$ is a poor approximation of π_i, the GREG estimator may suffer from a design-bias. On the other hand, it is model-unbiased and model-consistent for t_y, regardless of the quality of the approximation $\widehat{\pi}_i$, provided that Model (1.6) holds. Borrowing from the missing data literature, the GREG estimator is said to be doubly robust or doubly protected, as being design-consistent if the π_i's are known without errors even if (1.6) is misspecified, and being model-consistent if (1.6) holds even if the sampling design is misspecified and the π_i's are replaced by some poor approximation; e.g., Kott and Liao (2012) and Kim and Haziza (2014) for a discussion of doubly robust procedures in the context of finite population sampling.

For rejective sampling (Fuller (2009)), the inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)^\top$ are unknown and some approximations are needed. The π_i's can be approximated through Monte Carlo methods (see Section 3) or through Edgeworth expansions (see Section 4). Fuller (2009) suggests the use of a GREG-type estimator based on the basic inclusion probabilities p_i, which are different from the inclusion probabilities π_i. As a result, the sampling design is misspecified. We argue in Section 3 that, although the GREG-type estimator advocated by Fuller (2009) is design-consistent, it may suffer from bias for finite sample sizes when the p_i's do not provide a good approximation of the π_i's, unless (1.6) holds. In contrast, the Cube method ensures that the π_i's are exactly satisfied. As a result, the Horvitz-Thompson estimator $\widehat{t}_y^{\pi}$ is design-unbiased and design-consistent, although it may suffer from a greater variability. To control for the potential instability of the Horvitz-Thompson estimator, it is recommended, with the Cube method, to use a GREG estimator based on the true inclusion probabilities π_i and the vector of balancing variables. This is discussed in Section 2.

2. The Cube Algorithm

The cube method (Deville and Tillé (2004)) enables one to select (approximately) $\pi\mathbf{z}$-balanced samples such that the inclusion probabilities $\boldsymbol{\pi}$ are exactly respected. The cube method consists of two distinct steps: a flight phase, at the end of which an exact balancing is maintained, and a landing phase in which the balancing equations may be partly relaxed until the complete sample is obtained, while the inclusion probabilities remain exactly respected.

The flight phase (Deville and Tillé (2004); Chauvet and Tillé (2006); Tillé (2011)) is a random walk starting from the vector of inclusion probabilities $\boldsymbol{\pi}$ and ending at a random vector $\boldsymbol{\pi}^*$ such that $\pi_i^* = 0$ if unit i is definitely rejected from the sample, $\pi_i^* = 1$ if unit i is selected and $0 < \pi_i^* < 1$ if the decision for unit i remains pending at the end of the flight phase. Denote by P^* the set of units such that $0 < \pi_i^* < 1$, so that $I_i = \pi_i^*$ for $i \notin P^*$. From Proposition 1 in Deville and Tillé (2004), the size of P^* is at most the number J of balancing variables. The flight phase is performed in such a way that

$$E_F\left(\boldsymbol{\pi}^*\right) = \boldsymbol{\pi}, \tag{2.1}$$

$$\widehat{t}_{\mathbf{z}}^{\pi *} \equiv \sum_{i \in P} \frac{\mathbf{z}_i}{\pi_i} \pi_i^* = \sum_{i \in P} \mathbf{z}_i, \tag{2.2}$$

where the subscript F denotes the flight phase. Equation (2.1) ensures that the inclusion probabilities are exactly respected at the end of the flight phase, while (2.2) ensures that the pseudo-estimator $\widehat{t}_{\mathbf{z}}^{\pi *}$ is exactly $\pi\mathbf{z}$-balanced. The flight phase does not lead to an estimator per se, since the selection is still not carried through for the units in P^*.

The objective of the landing phase is to complete the sample selection process, either by successively relaxing the balancing equations or by means of an enumerative algorithm on P^* (Tillé (2011, p.163)). In any case, the landing phase is performed to obtain a vector of sample indicators $\mathbf{I}$ such that

$$E_L\left(\mathbf{I} \,|\boldsymbol{\pi}^*\right) = \boldsymbol{\pi}^*, \tag{2.3}$$

$$\widehat{t}_{\mathbf{z}}^{\pi} \equiv \sum_{i \in P} \frac{\mathbf{z}_i}{\pi_i} I_i \simeq \sum_{i \in U} \mathbf{z}_i, \tag{2.4}$$

where the subscript L denotes the landing phase. Equation (2.3) ensures that the inclusion probabilities are exactly respected at the end of the landing phase since, from (2.1) and (2.3),

$$E_p(\mathbf{I}) = E_F E_L(\mathbf{I}|\boldsymbol{\pi}^*) = E_F(\boldsymbol{\pi}^*) = \boldsymbol{\pi},$$

where the subscript p denotes the global sampling design. Consequently, the Horvitz-Thompson estimator $\widehat{t}_y^{\pi}$ is exactly design-unbiased for t_y.

The non-exact balancing (see (2.4)) results in an additional variability for the Horvitz-Thompson estimator, which is due to the landing phase. More precisely, the variance of $\widehat{t}_y^{\pi}$ can be written as

$$\begin{aligned} V_p(\widehat{t}_y^{\pi}) &= V_F\left\{E_L\left(\widehat{t}_y^{\pi}\,|\boldsymbol{\pi}^*\right)\right\} + E_F\left\{V_L\left(\widehat{t}_y^{\pi}\,|\boldsymbol{\pi}^*\right)\right\} \\ &= V_F\Big(\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^*\Big) + E_F\Big\{V_L\Big(\sum_{i\in P^*}\frac{y_i}{\pi_i}I_i\Big|\,\boldsymbol{\pi}^*\Big)\Big\}. \end{aligned} \tag{2.5}$$

The first term on the right-hand side of (2.5) is the variance due to the flight phase, whereas the second term is due to the landing phase. It follows from (2.2) that, for any q-vector $\mathbf{B}$, we have

$$\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^* = \sum_{i\in P}\frac{y_i-\mathbf{B}^\top\mathbf{z}_i}{\pi_i}\pi_i^* + \mathbf{B}^\top\sum_{i\in P}\mathbf{z}_i.$$

As a result, we can write

$$V_F(\widehat{t}_y^{\pi}) \equiv V_F\Big(\sum_{i\in P}\frac{y_i}{\pi_i}\pi_i^*\Big) = V_F\Big(\sum_{i\in P}\frac{E_i}{\pi_i}\pi_i^*\Big), \tag{2.6}$$

where $E_i = y_i - \mathbf{B}^\top\mathbf{z}_i$. The variance due to the landing phase is

$$V_L(\widehat{t}_y^{\pi}) \equiv E_F\Big\{V_L\Big(\sum_{i\in P^*}\frac{y_i}{\pi_i}I_i\Big|\,\boldsymbol{\pi}^*\Big)\Big\} = E_F\Big\{\sum_{i\in P^*}\sum_{j\in P^*}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}(\pi_{ij}^* - \pi_i^*\pi_j^*)\Big\}, \tag{2.7}$$

where $\pi_{ij}^* = E_L(I_iI_j|\boldsymbol{\pi}^*)$. If the balancing variables have a large explanatory power for the variable y, the residuals E_i are small and, from (2.6), so is the variance due to the flight phase. As a result, the contribution of the variance due the landing phase to the total variance may be appreciable (Breidt and Chauvet (2011)). Further, from (2.7), the variance due to the landing phase may be appreciable if the number of balancing variables is large as compared to the sample size (in which case, the random subpopulation P^* may be large as well) and/or the inclusion probabilities π_i are poorly or negatively related to the variable y, so that the y_i/π_i are highly variable (Chauvet (2011)).

The Horvitz-Thompson estimator $\widehat{t}_y^{\pi}$ is design-unbiased for t_y even if the balancing equations (1.2) are not satisfied because the inclusion probabilities are exactly satisfied. To cope with the rounding problem, it is recommended to perform some form of calibration on the set of balancing variables (Tillé (2011)). For example, one can use the GREG estimator, $\widehat{t}_{reg}^{\pi} = \sum_{i\in S} w_iy_i$, where the weights w_i are given by (1.5) with $\mathbf{u}_i = \mathbf{z}_i$. The error of $\widehat{t}_{reg}^{\pi}$ can be expressed as

$$\widehat{t}_{reg}^{\pi} - t_y = \Big(\sum_{i\in S}\pi_i^{-1}E_i^{\pi} - \sum_{i\in P}E_i^{\pi}\Big) + (\widehat{\mathbf{B}}^{\pi} - \mathbf{B}^{\pi})^\top(\mathbf{t}_{\mathbf{z}} - \widehat{\mathbf{t}}_{\mathbf{z}}^{\pi}), \tag{2.8}$$

where

$$\widehat{\mathbf{B}}^{\pi}=\Big(\sum_{i\in S}\pi_i^{-1}\mathbf{z}_i\mathbf{z}_i^{\top}\Big)^{-1}\sum_{i\in S}\pi_i^{-1}\mathbf{z}_iy_i,\qquad \mathbf{B}^{\pi}=\Big(\sum_{i\in P}\mathbf{z}_i\mathbf{z}_i^{\top}\Big)^{-1}\sum_{i\in P}\mathbf{z}_iy_i$$

and $E_i^{\pi}=y_i-\mathbf{z}_i^{\top}\mathbf{B}^{\pi}$. Noting that $E_p\left(\sum_{i\in S}\pi_i^{-1}E_i^{\pi}-\sum_{i\in P}E_i^{\pi}\right)=0$, the design-bias of $\widehat{t}_{reg}^{\pi}$ is given by

$$E_p(\widehat{t}_{reg}^{\pi}-t_y)=E_p\left\{(\widehat{\mathbf{B}}^{\pi}-\mathbf{B}^{\pi})^{\top}(\mathbf{t}_{\mathbf{z}}-\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})\right\}. \tag{2.9}$$

The expectation in (2.9) is often referred to as the small sample bias. Therefore if the sample size n is not large enough, the small sample bias may be significant. Under mild regularity conditions (Deville and Särndal (1992)), we have $\widehat{\mathbf{B}}-\mathbf{B}^{\pi}=O_p(n^{-1/2})$. For non-balanced sampling designs (e.g., simple random sampling without replacement), the term $(\mathbf{t}_{\mathbf{z}}-\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})$ is $O_p(Nn^{-1/2})$. As a result, $(\widehat{\mathbf{B}}^{\pi}-\mathbf{B}^{\pi})^{\top}(\mathbf{t}_{\mathbf{z}}-\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})=O_p(Nn^{-1})$. On the other hand, with the Cube method, Deville and Tillé (2004) showed that $(\mathbf{t}_{\mathbf{z}}-\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})=O_p(NJn^{-1})$, which implies that $(\widehat{\mathbf{B}}^{\pi}-\mathbf{B}^{\pi})^{\top}(\mathbf{t}_{\mathbf{z}}-\widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})=O_p(NJn^{-3/2})$. If $J=O(1)$, this is $O_p(Nn^{-3/2})$. Therefore, the strategy consisting of the Cube method and the GREG estimator is less vulnerable to small sample bias than the strategy consisting of a non-balanced sampling design, such as simple random sampling without replacement and the GREG estimator.

3. Rejective Sampling

In this section, we adopt the notation of Fuller (2009) who studied the properties of rejective sampling procedures that consists of discarding any sample that does not meet a specified balancing tolerance. Samples are selected using a basic procedure based on the vector of design variables, $\mathbf{z}$, available for all $i\in P$. Commonly used procedures are simple random sampling without replacement, stratified sampling, and Poisson sampling. A subset of the $\mathbf{z}$-variables are the $\mathbf{z}_2$-variables, which are those satisfying $V_b(\bar{\mathbf{z}}_2^p)=\mathbf{0}$, where $\bar{\mathbf{z}}_2^p=N^{-1}\sum_{i\in S_b}p_i^{-1}\mathbf{z}_{2i}$ is the vector of basic expansion estimators for the population mean of $\mathbf{z}_2$. The design variables not included in the set of the $\mathbf{z}_2$-variables are referred to as the $\mathbf{z}_1$-variables. Let $\mathbf{x}_i=\mathbf{z}_{1i}-\mathbf{C}^{\top}\mathbf{z}_{2i}$, where $\mathbf{C}$ is the matrix that minimizes $\mathrm{tr}\left\{V_b(\bar{\mathbf{z}}_1^p-\mathbf{C}^{\top}\bar{\mathbf{z}}_2^p)\right\}$ and $\bar{\mathbf{z}}_1^p=N^{-1}\sum_{i\in S_b}p_i^{-1}\mathbf{z}_{1i}$. Finally, let $\mathbf{z}=(\mathbf{x}^{\top},\mathbf{z}_2^{\top})^{\top}$.

The rejective procedure $p(\cdot)$ proceeds as follows: (i) select a random sample, S_b, according to the basic procedure, using $p_1,\ldots,p_N$ as the vector of inclusion probabilities; (ii) the sample is retained if the rejection rule (1.3) is satisfied, with $\mathbf{z}$ replaced by $\mathbf{x}$, and with $\gamma>0$ a constant specified by the survey statistician; otherwise, replace the sample in the population and repeat step (i). We denote by S the final random sample.

A small value of γ corresponds to a high rejection rate. The π_i's are complex functions of $\widehat{\mathbf{t}}_{\mathbf{x}}^{\pi}$, $\mathbf{t}_{\mathbf{x}}$, and the p_i's; as a result, they are generally untractable.

Although (1.3) ensures that the sample S is approximately $p\mathbf{x}$-balanced for small values of γ, there is no guarantee that it is $\pi\mathbf{x}$-balanced. Since the π_i's are unknown, they must be approximated. Some approximations are discussed in Sections 3.1-3.3, and in Section 4.

3.1. Basic estimator

We start by examining the bias of the expansion estimator

$$\widehat{t}_y^p = \sum_{i \in S} p_i^{-1} y_i \tag{3.1}$$

based on the basic inclusion probabilities p_i. Although $\widehat{t}_y^p$ is design-unbiased for t_y with respect to the basic procedure, it is generally biased with respect to the rejective sampling procedure. The bias is given by

$$B_p(\widehat{t}_y^p) = \sum_{i \in P} \delta_i y_i, \qquad \text{where} \qquad \delta_i = \frac{\pi_i - p_i}{p_i}, \tag{3.2}$$

with the subscript p denoting the expectation with respect to the rejective sampling design. The coefficient $|\delta_i|$ can be viewed as a measure of relative distance between the basic inclusion probabilities, p_i, and the inclusion probabilities with respect to the rejective sampling procedure, π_i. The bias in (3.2) is large if some units exhibit a large y-value and/or a large δ-value. For simplicity, consider the case of a scalar x and suppose that the basic procedure is simple random sampling without replacement so that $p_i = n/N$ for all $i \in P$. We expect a large value of $|\delta_i|$ if x_i is much larger than the population mean $\bar{X} = t_x/N$. In this case, π_i is expected to be significantly smaller than p_i because most samples containing unit i are likely to be rejected, which in turns, leads to a large value of $|\delta_i|$. Therefore, a unit exhibiting large values of both x and y may contribute significantly to the bias of $\widehat{t}_y^p$. To overcome this problem, it may be wise to construct an additional stratum consisting of all the units exhibiting large x-values. However, in practice, strata are usually formed for operational convenience. As a result, it is not unusual for some strata to include units with relatively large x-values. As the sample size n increases, we expect the δ-values to become smaller. Therefore, we expect the bias of the basic expansion estimator, $\widehat{t}_y^p$, to decrease as the sample size increases. This is confirmed by the empirical results presented in Section 5.

3.2. Estimator based on Monte Carlo approximations

One option consists of estimating the π_i's through Monte Carlo simulations to obtain $\widehat{\pi}_i^{MC}$; see Fattorini (2006), Thompson and Wu (2008), and Lesage (2013). Then, use the Monte Carlo expansion estimator

$$\widehat{t}_y^{\pi} = \sum_{i \in S} \frac{y_i}{\widehat{\pi}_i^{MC}} \tag{3.3}$$

as an estimator of t_y. For large populations though, simulating enough samples to obtain precise estimates of the inclusion probabilities may prove problematic. Alternatively, one may use a GREG estimator based on the $\widehat{\pi}_i^{MC}$'s, given by (1.4), where w_i is given by (1.5) with π_i replaced by $\widehat{\pi}_i^{MC}$.

3.3. The regression estimator

An alternative option was studied in Fuller (2009), who showed that the regression estimator based on the basic inclusion probabilities p_i and the vector of auxiliary variables $\mathbf{z}$, is design-consistent. Fuller (2009) assumes that $p_b(\cdot)$ is such that there exists some ϕ_i satisfying

$$E(\bar{\mathbf{x}}^p - \bar{\mathbf{X}}|i \in S_b) = p_i^{-1} N^{-1} \phi_i \mathbf{x}_i, \tag{3.4}$$

where $\bar{\mathbf{x}}^p = N^{-1}\sum_{i\in s_b} p_i^{-1}\mathbf{x}_i$ and $\bar{\mathbf{X}} = N^{-1}\sum_{i\in P}\mathbf{x}_i$. Fuller (2009) advocated the use of the GREG type estimator

$$\widehat{t}^p_{\text{reg}} = \sum_{i\in P} \mathbf{z}_i^\top \widehat{\mathbf{B}}^p \text{ with } \widehat{\mathbf{B}}^p = \Big(\sum_{i\in S} p_i^{-2}\phi_i \mathbf{z}_i \mathbf{z}_i^\top\Big)^{-1} \sum_{i\in S} p_i^{-2}\phi_i \mathbf{z}_i y_i. \tag{3.5}$$

Note that for an arbitrary basic procedure,

$$E(\bar{\mathbf{x}}^p - \bar{\mathbf{X}}|i \in S_b) = N^{-1}\sum_{j\in U}\left(\frac{p_{ij} - p_i p_j}{p_i p_j}\right)\mathbf{x}_j,$$

where $p_{ij} = pr(i, j \in S_b)$, so that (3.4) may not be fulfilled for any sampling design $p_b(\cdot)$. Fuller (2009) showed that the estimator (3.5) is design-consistent provided there exists a vector of constants $\boldsymbol{\lambda}$ such that

$$p_i^{-2}\phi_i \mathbf{z}_i^\top \boldsymbol{\lambda} = p_i^{-1}. \tag{3.6}$$

Example 1. Suppose that an auxiliary variable z_{1i} is available. If the basic procedure is simple random sampling without replacement with basic inclusion probabilities $p_i = n/N$ for all i, we have $\mathbf{z}_{2i} = 1$ and $\mathbf{z}_{1i} = z_{1i}$. Also, $\mathbf{z}_i^\top = (\mathbf{x}_i^\top, 1)$ with $\mathbf{x}_i = z_{1i} - \bar{Z}_1$ and $\bar{Z}_1 = N^{-1}\sum_{i\in P} z_{1i}$. Since

$$E(\bar{\mathbf{x}}^p - \bar{\mathbf{X}}|i \in S_b) = (N-1)^{-1}\left(\frac{N}{n} - 1\right)\mathbf{x}_i,$$

we have $\phi_i = (N-1)^{-1}(N-n)$ which does not depend on i.

Example 2. Suppose that an auxiliary variable z_{1i} is available. If the basic procedure is Bernoulli sampling with basic inclusion probabilities p_0 for all i, we have $\mathbf{z}_{2i} = \emptyset$, $\mathbf{z}_{1i} = (1, z_{1i})^\top$, and $\mathbf{z}_i^\top = (\mathbf{x}_i^\top, \mathbf{z}_{2i}^\top)$ with $\mathbf{x}_i = (1, z_{1i})^\top$. Since

$$E(\bar{\mathbf{x}}^p - \bar{\mathbf{X}}|i \in S_b) = N^{-1}\frac{(1-p_0)}{p_0}\mathbf{x}_i,$$

we have $\phi_i = 1 - p_0$ which does not depend on i.

The regression estimator (3.5) can be written in two alternative forms. Using (3.6), it can expressed as

$$\widehat{t}^p_{\mathrm{reg}} = \sum_{i\in S} \frac{y_i}{\widehat{\pi}_i^F}, \tag{3.7}$$

where

$$\widehat{\pi}_i^F = p_i\Big[\Big\{1 + (t_{\mathbf{z}} - \widehat{t}^p_{\mathbf{z}})^\top \Big(\sum_{i\in S} p_i^{-2}\phi_i \mathbf{z}_i\mathbf{z}_i^\top\Big)^{-1} p_i^{-2}\phi_i\mathbf{z}_i\Big\}\Big]^{-1}. \tag{3.8}$$

Thus, $\widehat{\pi}_i^F$ can be viewed as an implicit estimate of the true inclusion probability π_i. As the sample size increases, the difference between p_i and $\widehat{\pi}_i^F$ becomes smaller. For finite sample sizes, this approximation may not be appropriate, especially if some units exhibit large values for the balancing variables. This is illustrated empirically in Section 5. A better approximation of the true π_i's can be obtained through Edgeworth expansions. This is discussed in Section 4 with Poisson sampling as the basic procedure.

Alternatively, the regression estimator (3.5) can be written as

$$\widehat{t}^p_{\mathrm{reg}} = \widehat{t}^p_y + (\mathbf{t_z} - \widehat{\mathbf{t}}^p_{\mathbf{z}})^\top \widehat{\mathbf{B}}^p. \tag{3.9}$$

From (3.9), it follows that $\widehat{t}^p_{\mathrm{reg}}$ and $\widehat{t}^p_y$ are expected to share similar properties in terms of bias and efficiency for small values of the balancing tolerance γ, which correspond to high rejection rates. This is due to the fact that a high rejection rate corresponds to small values of $\mathbf{t_z} - \widehat{\mathbf{t}}^p_{\mathbf{z}}$.

We now turn to the bias of $\widehat{t}^p_{\mathrm{reg}}$, whose error can be expressed as

$$\widehat{t}^p_{\mathrm{reg}} - t_y = (\widehat{t}^p_E - t_E) + (\widehat{\mathbf{B}}^p - \mathbf{B}^p)^\top (\mathbf{t_z} - \widehat{\mathbf{t}}^p_{\mathbf{z}}), \tag{3.10}$$

where

$$\mathbf{B}^p = \Big(\sum_{i\in P} \pi_i p_i^{-2}\phi_i\mathbf{z}_i\mathbf{z}_i^\top\Big)^{-1} \sum_{i\in P} \pi_i p_i^{-2}\phi_i\mathbf{z}_i y_i$$

and $\widehat{t}^p_E = \sum_{i\in S} p_i^{-1}E_i$, $t_E = \sum_{i\in P} E_i$ with $E_i = y_i - \mathbf{z}_i^\top \mathbf{B}^p$. From (3.10), the design-bias of $\widehat{t}^p_{\mathrm{reg}}$ can be expressed as the sum of two terms.

The design-expectation of the first term on the right hand-side of (3.10) is

$$E_p(\widehat{t}^p_E - t_E) = \sum_{i\in P} \delta_i E_i = \sum_{i\in P} (\delta_i - \bar{\delta})(E_i - \bar{E}), \tag{3.11}$$

where $\bar{\delta} = N^{-1}\sum_{i\in P}\delta_i$, noting that $\bar{E} = N^{-1}\sum_{i\in P} E_i = 0$. The design-expectation of $\widehat{t}^p_E - t_E$ is zero if the E_i's and the δ_i's are unrelated. This condition is approximately satisfied if

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \text{ with } E_m(\epsilon_i|\mathbf{x}_i) = 0. \tag{3.12}$$

On the other hand, the term (3.11) may be large if some units exhibit a large E-value and/or a large δ-value, which can occur, for example, when the relationship between y and $\mathbf{z}$ is not linear. In multipurpose surveys, it is unrealistic to presume that (3.12) holds for every characteristic of interest y, in which case the term $E_p(\widehat{t}_E^p - t_E)$ may be significantly different from zero.

The design-expectation of the second term on the right hand-side of (3.10) is $E_p\left\{(\widehat{\mathbf{B}}^p - \mathbf{B}^p)^\top(\mathbf{t_z} - \widehat{\mathbf{t}}_{\mathbf{z}}^p)\right\}$, referred to as the small sample bias. Assuming that $\gamma = O(1)$, $\mathbf{t_z} - \widehat{\mathbf{t}}_{\mathbf{z}}^p$ is $O_p(Nn^{-1/2})$, so that the second term on the right hand-side of (3.10) is $O_p(Nn^{-1})$. Therefore, if the sample size n is not large enough, the small sample bias may be significant.

In summary, under Fuller's estimation procedure, there are two possible sources of bias: the first is due to a misspecification of the sampling design, based on the basic inclusion probabilities p_i instead of the true inclusion probabilities π_i. The second is the small sample bias that can be large for finite sample sizes. Both sources of bias are illustrated numerically in Section 1 of the Supplementary Material.

4. Approximation of the Inclusion Probabilities through Edgeworth Expansion

In this section, we obtain an approximation of the true inclusion probabilities π_i through Edgeworth expansion when Poisson sampling is used as the basic procedure. For simplicity, we consider the case of a scalar x. Let $\mathbf{I}_b = (I_{b1}, \ldots, I_{bN})^\top$ have entries $I_{bi} = 1$ if unit i is selected in the basic random sample S_b and $I_{bi} = 0$, otherwise. Condition (1.3) may be rewritten as $-\gamma \le X \le \gamma$ with $X = (1/\sqrt{d}) \sum_{i \in P} \tilde{x}_i (I_i - p_i)$, $d = \sum_{i \in P} \tilde{x}_i^2 \, p_i(1-p_i)$ and $\tilde{x}_i = x_i/p_i$. The final inclusion probability for rejective sampling is

$$\pi_i = p_i \frac{pr(-\gamma \le X \le \gamma | I_{bi} = 1)}{pr(-\gamma \le X \le \gamma)}. \tag{4.1}$$

We prove in Section 3 of the Supplementary Material that

$$\pi_i = p_i \left\{ 1 - \frac{1}{d} \frac{\gamma\phi(\gamma)}{2\psi(\gamma) - 1} \tilde{x}_i^2 (1-p_i)(1-2p_i) + \frac{\kappa_3}{\sqrt{d}} \frac{\gamma\phi(\gamma)(3-\gamma^2)}{3(2\psi(\gamma)-1)} \tilde{x}_i (1-p_i) \right\} \\ +o(d^{-1}). \tag{4.2}$$

Assuming that $\gamma = O(d^{-0.5})$, this approximation simplifies to

$$\pi_i = p_i \left\{ 1 - \frac{1}{2d} \tilde{x}_i^2 (1-p_i)(1-2p_i) + \frac{\kappa_3}{2\sqrt{d}} \tilde{x}_i (1-p_i) \right\} + o(d^{-1}). \tag{4.3}$$

Ignoring the $o(d^{-1})$ term in (4.3), we obtain an alternative approximation of π_i denoted $\widehat{\pi}_i^{ED}$. An Edgeworth expansion estimator of t_y is given by

$$\widehat{t}_y^{ED} = \sum_{i\in S} \frac{y_i}{\widehat{\pi}_i^{ED}}. \tag{4.4}$$

A comparison of (3.8) and (4.3) helps in understanding why the GREG type estimator (3.5) advocated by Fuller (2009), can perform poorly in terms of bias in finite samples. Both (3.8) and (4.3) exhibit significant differences and as a result, (3.8) may not be a good approximation of π_i, unlike (4.3). This is illustrated empirically in Section 5.1. An alternative to (4.4) is a GREG estimator based on the $\widehat{\pi}_i^{ED}$'s, given by (1.4), where w_i is given by (1.5) with π_i replaced by $\widehat{\pi}_i^{ED}$.

5. Simulation Study

We conducted an extensive simulation study in order to compare several approximations of the first-order inclusion probabilities, and to compare the performance of several estimators in terms of relative bias and mean square error. We generated 18 populations of size $N = 500$, each consisting of an auxiliary variable x and a characteristic of interest y. In each population, the x-values were first generated according to the distributions: a normal with mean 2 and variance 1; a mixture, where 99% were generated from a normal with mean 2 and variance 1 and the remaining observations were set manually to $7.9, 8.0, 8.2, 8.3$, significantly larger than the remaining observations; a log-normal with mean 0 and variance 0.9.

Given the x-values, the y-values were generated according to the models: $y_i = 1+2(x_i-\bar{X})+\sigma\ \varepsilon_i$ linear; $y_i = 1+2(x_i-\bar{X})^2+\sigma\ \varepsilon_i$, quadratic; $y_i = \exp\{1+1(x_i-\bar{X})\}+\sigma\ \varepsilon_i$, exponential; $y_i = 1+2(x_i-\bar{X})^2 - 10\exp\left\{-20(x_i-\bar{X})^2\right\} + \sigma\ \varepsilon_i$, bump; $y_i = 20I(x_i \in [1.077;\ 7.66]) + \varepsilon_i$, anova; $y_i \sim \mathcal{B}(1,\Phi_i)$, where $\log\left(\Phi_i/(1-\Phi_i)\right) = 2\ (x_i - 2)$, logistic. For each model, we used $\sigma = 1$ and the errors ε_i were generated from a normal with mean 0 and variance 1. Table 2 of the Supplementary Material presents several characteristics for these populations.

5.1. Approximations of the first-order inclusion probabilities

We compared several approximations of the first-order inclusion probabilities when the samples were selected according to the rejective procedure of Fuller (2009) with Bernoulli sampling as the basic procedure. The expected sample size n was set to $25, 50$, and 100. Samples were repeatedly selected until the rejection rule (1.3) was satisfied, with $\mathbf{z}_i = x_i$. The balancing tolerance γ was set so that approximately 90% of the samples were rejected.

In each sample, we computed the basic selection probabilities p_i (Basic); the estimated probability $\widehat{\pi}_i^F$ (Fuller) given in (3.8); the Monte-Carlo approximation $\widehat{\pi}_i^{MC}$ (MC) used in equation (3.3); the approximation $\widehat{\pi}_i^{ED}$ (Edge.) obtained in (4.3) through Edgeworth expansions. The $\widehat{\pi}_i^{MC}$'s were obtained through an independent set of $K_1 = 500{,}000$ simulations.

Figure 1 plots the relationship between the different inclusion probabilities and the x-variable. In each plot, the bold dashed line corresponds to the basic inclusion probabilities p_i, that are all equal under Bernoulli sampling; the bold black curve represents the Monte Carlo average of $\widehat{\pi}_i^F$; the dashed curve represent the Monte Carlo approximation, $\widehat{\pi}_i^{MC}$; the black curve represent the approximation of the π_i's obtained through Edgeworth expansions, $\widehat{\pi}_i^{ED}$. The blue curve is a smoothed adjustment curve.

From Figure 1, when the x-values were generated from a normal distribution, the probabilities $\widehat{\pi}_i^F$, $\widehat{\pi}_i^{MC}$ and $\widehat{\pi}_i^{ED}$ were relatively close to the basic inclusion probabilities p_i. This was especially true for $n = 50$ and $n = 100$. When the x-values were generated according to a mixture or a lognormal distribution, the units with a large x-value exhibited an inclusion probability $\widehat{\pi}_i^{MC}$ significantly smaller than the basic inclusion probability p_i. This was especially apparent for $n = 25$, where the probabilities $\widehat{\pi}_i^F$ given by (3.8) provided a poor approximation of the true π_i's, especially for units associated with large x-values. We note that the distribution of the $\widehat{\pi}_i^{ED}$'s was very close to that of the $\widehat{\pi}_i^{MC}$'s. Finally, as the (expected) sample size increased, all the approximations of the π_i's became increasingly closer to the basic inclusion probabilities p_i, as expected.

5.2. Efficiency of sampling and estimation strategies

We compared several sampling and estimation strategies in terms of relative bias and relative efficiency. From each population, we selected $K_2 = 10{,}000$ samples of size $n = 25,\ 50$ and 100, according to the sampling designs: simple random sampling without replacement; the rejective procedure of Fuller (2009) described in Section 3 with simple random sampling without replacement as the basic sampling procedure; the basic samples were selected until $\left|(\hat{t}_x^p - t_x)/V_b(\hat{t}_x^p)^{1/2}\right| < \gamma$, and the balancing tolerance γ was set so that the rejection rate was approximately equal to 90% and 50%; the cube method described in Section 2 with the balancing constraints of fixed sample size and $\hat{t}_x^\pi = t_x$ with $\pi_i = n/N$.

We were interested in estimating the population total of the y-values, $t_y = \sum_{i \in P} y_i$. In each sample selected by the rejective procedure, we computed the estimator given by (3.1)(Basic); the estimator given by (3.5), where the values of p_i, ϕ_i and $\mathbf{z}_i$ are presented in Example 1(Fuller); the estimator given by (3.3) (MC); the estimator given by (1.4), where w_i is given by (1.5) with $\mathbf{u}_i = (1, x_i)^\top$ and π_i replaced by $\widehat{\pi}_i^{MC}$ (MC-Reg).

In each sample selected by the Cube method, we computed estimator given by (1.4) with $w_i = \pi_i^{-1}$ (Cube); the estimator given by (1.4), where w_i is given by (1.5) with $\mathbf{u}_i = (1, x_i)^\top$ (Cube-Reg). In each sample selected by simple random sampling without replacement, we computed the estimator given by (1.4) with

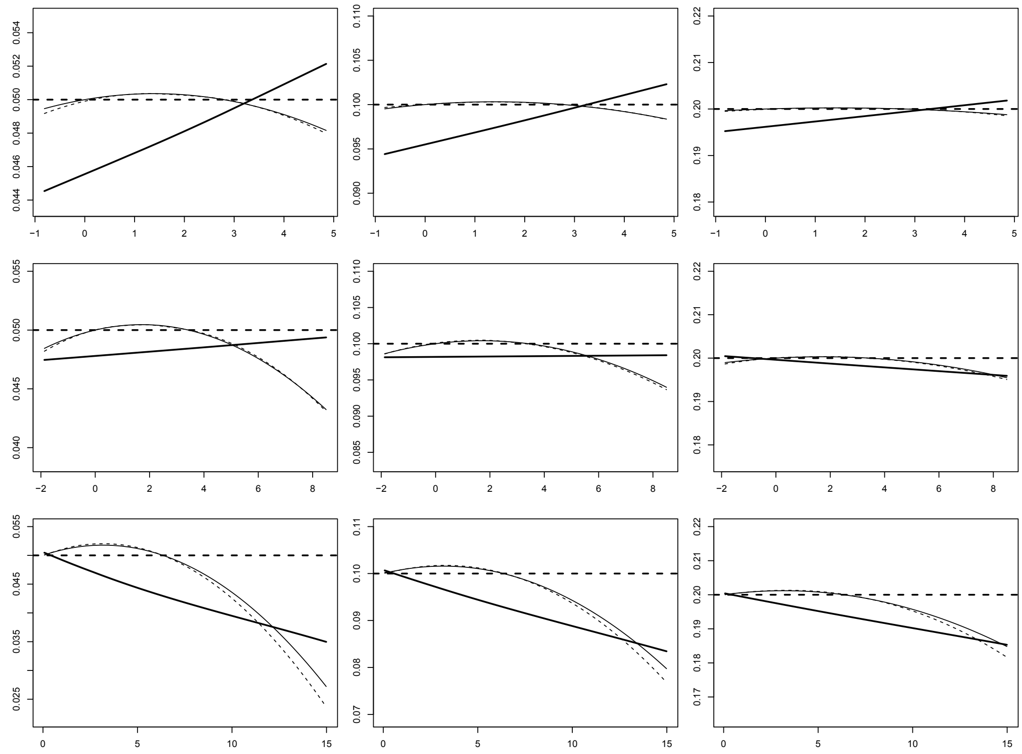

Figure 1. Smoothed curve of the inverse of the weights of BEE (in bold dashed), MC (in dashed), Fuller (in bold black) and Edge. (in black), for a normal distribution on top, a mixture distribution in the middle and a log-normal distribution on the bottom, for a sample of size $n = 25$ on the right, $n = 50$ in the center and $n = 100$ on the left, for a rejective Bernoulli sampling, with an 90 % rejection rate.

$w_i = N/n$ (SRS); the estimator given by (1.4), where w_i is given by (1.5) with $\mathbf{u}_i = (1, x_i)^\top$ and $\pi_i = n/N$ (SRS-Reg).

As a measure of the bias of an estimator $\hat{t}$, we computed its Monte Carlo percent relative bias (RB)

$$RB_{MC}\left(\widehat{t}\right) = \frac{1}{K_2} \sum_{j=1}^{K_2} \frac{\left(\widehat{t^{(j)}} - t_y\right)}{t_y} \times 100,$$

where $\widehat{t}^{(j)}$ denotes the estimator $\widehat{t}$ for the j-th iteration, $j = 1, \ldots, K_2$. As a measure of variability of $\widehat{t}$, we computed its Monte Carlo percent coefficient of variation (CV)

$$CV_{MC}(\widehat{t}) = 100 \times \frac{\left\{(1/K_2)\sum_{j=1}^{K_2}\left(\widehat{t}^{(j)} - (1/K_2)\sum_{k=1}^{K_2}\widehat{t}^{(k)}\right)^2\right\}^{1/2}}{t_y}.$$

As a measure of relative efficiency (RE) of $\widehat{t}$, using the estimator $\widehat{t}^{p}_{reg}$ advocated by Fuller (2009) as the reference, we computed

$$RE_{MC}(\widehat{t}) = 100 \times \frac{\left\{(1/K_2)\sum_{j=1}^{K_2}\left(\widehat{t}^{p(j)}_{reg} - t_y\right)^2\right\}^{1/2}}{\left\{(1/K_2)\sum_{j=1}^{K_2}\left(\widehat{t}^{(j)} - t_y\right)^2\right\}^{1/2}}.$$

Tables 1−9 show the Monte Carlo results of eight estimators in terms of relative bias, coefficient of variation, and relative root mean square error for eighteen populations, and a rejection rate of 90%. The results corresponding to a rejection rate of 50% are presented in Section 2 of the Supplementary Material.

When the x-values were normally distributed, both Basic and Fuller showed small biases, regardless of the sample size n and the model used to generate the y-values; see Tables 1, 4, and 7. When the relationship between y and x was linear, Basic and Fuller showed virtually no bias, regardless of the distribution of the x-variable. They exhibited almost identical efficiency. When the x-variable was not normally distributed, Basic and Fuller were generally biased and their bias was virtually identical. The bias was especially large for highly non-linear relationships between y and x. For example, when the distribution of the x-values was log-normal and the relationship between y and x was exponential, both Basic and Fuller showed a value of RB approximately equal to −35%, see Table 6. In terms of efficiency, Basic was very close to Fuller in all the scenarios, with a value of RE ranging from 94% to 102%. The fact that Basic and Fuller exhibited almost identical properties in all the scenarios can be easily explained by the fact that the term $\mathbf{t_z} - \widehat{\mathbf{t}}^{p}_{\mathbf{z}}$ on the right hand-side of (3.9) was close to zero due to the high rejection rate. As a result, Fuller essentially reduced to Basic. Finally, we note that the bias of Basic and Fuller decreased as the sample size increased, as expected.

We now turn to the estimators MC and MC-Reg. First, MC showed a small bias in all the scenarios, as expected. However, in some scenarios, it was considerably less efficient than Fuller. For example, when the distribution of the x-variable was a mixture, the values of RE was 46% for the quadratic relationship between y and x, and for $n = 25$, see Table 2. For both the ANOVA and the logistic populations, MC was slightly more efficient than Fuller in all the scenarios. Except for these populations, MC-reg performed better than MC in terms of RE, although the difference became smaller as the sample size increased. When the distribution of the x-values was not normal and the relationship between y

Table 1. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 90% and for a normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
	RB	-0	-0	0	0	-0	0	-0	-0
Linear	CV	19	19	19	19	20	19	42	19
	RE	98	100	97	100	86	103	20	96
	RB	-2	-2	-0	-0	0	-0	0	-4
Quadratic	CV	18	18	19	18	18	18	18	18
	RE	100	100	89	94	95	95	101	87
	RB	-1	-1	-0	-0	0	-0	0	-3
Exponential	CV	13	13	14	14	14	13	23	13
	RE	99	100	81	91	84	94	32	90
	RB	-5	-5	0	-0	0	-1	0	-11
Bump	CV	55	55	56	55	56	56	54	58
	RE	100	100	97	98	96	95	104	88
	RB	1	1	0	0	-0	0	-0	1
Anova	CV	7	7	7	7	7	7	10	7
	RE	100	100	108	100	94	97	52	89
	RB	-0	-0	-0	-0	0	0	0	-0
Logistic	CV	16	16	16	16	16	16	20	16
	RE	100	100	102	102	98	100	62	94

Table 2. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 90% and for a mixture distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
	RB	0	0	0	0	0	0	-0	-0
Linear	CV	21	21	24	21	26	21	53	21
	RE	98	100	73	100	65	100	15	96
	RB	-9	-9	0	-5	-0	-3	1	-13
Quadratic	CV	33	33	51	41	39	36	40	32
	RE	100	100	46	71	79	90	75	98
	RB	-27	-28	1	-16	-1	-11	2	-38
Exponential	CV	112	111	183	138	135	120	152	102
	RE	100	100	39	68	72	90	57	111
	RB	-15	-15	-0	-8	-0	-5	1	-22
Bump	CV	58	58	83	69	66	62	67	58
	RE	100	100	53	76	84	93	81	94
	RB	1	1	0	1	0	0	-0	2
Anova	CV	8	8	8	9	8	8	10	8
	RE	100	100	106	86	95	95	75	91
	RB	1	1	0	1	0	0	0	1
Logistic	CV	15	15	15	15	15	15	19	16
	RE	100	100	107	99	100	101	67	95

Table 3. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 25$, with a rejection rate equal to 90% and for a log-normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	-0	-0	-0	-0	0	0	0	0
	CV	19	19	54	19	33	19	66	19
	RE	94	100	12	100	32	101	8	92
Quadratic	RB	-15	-15	0	-12	-0	-9	1	-21
	CV	37	37	128	40	61	41	75	38
	RE	99	100	9	90	42	89	27	84
Exponential	RB	-74	-74	4	-71	1	-38	5	-67
	CV	165	164	703	179	321	208	336	152
	RE	99	100	7	87	31	72	29	117
Bump	RB	-19	-19	0	-15	-0	-11	2	-26
	CV	47	47	156	51	75	52	93	49
	RE	99	100	10	90	44	89	29	81
Anova	RB	4	4	-0	3	0	3	-1	8
	CV	19	19	18	20	18	19	21	22
	RE	100	100	109	95	108	95	81	68
Logistic	RB	3	3	0	2	-0	2	-0	5
	CV	24	24	23	24	25	25	30	26
	RE	100	100	110	100	100	98	68	84

Table 4. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 90% and for a normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	-0	-0	-0	-0	0	0	0	-0
	CV	13	13	13	13	13	13	29	13
	RE	98	100	98	100	93	100	20	100
Quadratic	RB	-1	-1	0	-0	0	-0	-0	-2
	CV	12	12	12	12	13	13	12	13
	RE	100	100	95	97	93	93	101	91
Exponential	RB	-1	-1	0	0	0	0	-0	-1
	CV	9	9	10	9	10	9	15	9
	RE	99	100	91	96	93	97	35	97
Bump	RB	-2	-2	0	0	0	0	-1	-6
	CV	38	38	38	38	38	38	38	39
	RE	100	100	99	99	96	96	99	90
Anova	RB	0	0	-0	-0	-0	-0	0	1
	CV	5	5	5	5	5	5	7	5
	RE	100	100	103	100	96	98	51	93
Logistic	RB	-0	-0	-0	-0	0	0	-0	-0
	CV	11	11	11	11	11	11	14	11
	RE	100	100	101	101	97	98	61	95

Table 5. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 90% and for a mixture distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	-0	-0	-0	-0	-0	-0	-0	-0
	CV	14	14	15	14	16	14	36	15
	RE	97	100	93	100	78	99	16	95
Quadratic	RB	-4	-4	-0	-2	0	-1	-0	-8
	CV	25	25	29	28	26	26	27	25
	RE	100	100	74	82	91	95	86	94
Exponential	RB	-14	-14	-2	-6	1	-2	0	-24
	CV	84	84	103	96	93	89	103	82
	RE	100	100	69	80	84	91	69	101
Bump	RB	-7	-7	-1	-2	1	-1	-0	-13
	CV	42	42	48	47	45	44	45	42
	RE	100	100	78	84	92	95	89	93
Anova	RB	1	1	0	0	-0	0	0	1
	CV	6	6	6	6	6	6	7	6
	RE	100	100	102	91	97	95	77	91
Logistic	RB	0	0	-0	0	-0	0	0	1
	CV	11	11	10	11	11	11	13	11
	RE	100	100	103	99	101	101	71	99

Table 6. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 50$, with a rejection rate equal to 90% and for a log-normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	-0	-0	-0	-0	0	-0	-0	-0
	CV	13	13	16	13	19	13	45	13
	RE	95	100	62	100	47	100	8	97
Quadratic	RB	-7	-7	-0	-4	-0	-3	0	-12
	CV	31	31	48	35	38	32	52	30
	RE	99	100	44	83	70	102	38	96
Exponential	RB	-35	-35	-0	-26	-0	-17	2	-46
	CV	178	178	277	203	217	185	228	150
	RE	99	100	43	78	70	95	63	133
Bump	RB	-8	-8	-0	-5	-0	-4	0	-15
	CV	39	39	59	44	47	40	64	38
	RE	99	100	46	84	72	102	40	94
Anova	RB	2	2	0	1	0	1	-0	4
	CV	13	13	13	14	13	13	14	14
	RE	100	100	105	95	109	102	87	82
Logistic	RB	1	1	0	1	0	1	0	3
	CV	17	17	16	17	17	17	21	18
	RE	100	100	107	99	102	101	67	91

Table 7. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 90% and for a normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	0	0	0	0	-0	-0	0	0
	CV	9	9	9	9	9	9	19	9
	RE	97	100	97	100	94	99	19	97
Quadratic	RB	-0	-0	-0	-0	-0	-0	0	-1
	CV	8	8	8	8	8	8	8	8
	RE	100	100	98	99	98	98	101	96
Exponential	RB	-0	-0	-0	-0	-0	0	0	-1
	CV	6	6	6	6	6	6	10	6
	RE	99	100	96	98	96	99	34	100
Bump	RB	-1	-1	0	0	0	-0	0	-2
	CV	25	25	25	25	25	25	25	26
	RE	100	100	100	100	99	99	97	94
Anova	RB	0	0	-0	-0	0	0	0	0
	CV	3	3	3	3	3	3	5	3
	RE	100	100	101	100	95	96	49	94
Logistic	RB	0	0	0	0	0	0	0	-0
	CV	7	7	7	7	7	7	9	7
	RE	100	100	100	100	98	99	61	99

Table 8. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 90% and for a mixture distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
Linear	RB	-0	0	-0	0	-0	-0	0	0
	CV	10	10	10	10	10	10	24	10
	RE	98	100	97	100	86	98	15	99
Quadratic	RB	-2	-2	-0	-0	-0	-0	0	-4
	CV	17	17	18	18	17	17	18	17
	RE	100	100	90	92	99	100	86	92
Exponential	RB	-5	-5	-0	-1	-0	-1	0	-12
	CV	59	58	63	62	60	59	69	59
	RE	100	100	88	91	96	99	73	95
Bump	RB	-3	-3	-0	-0	-0	-0	-0	-6
	CV	29	29	30	30	29	29	31	30
	RE	100	100	92	93	98	100	88	91
Anova	RB	0	0	0	0	-0	-0	0	1
	CV	4	4	4	4	4	4	4	4
	RE	100	100	101	97	100	99	77	91
Logistic	RB	0	0	0	0	-0	0	-0	0
	CV	7	7	7	7	7	7	8	7
	RE	100	100	101	100	97	97	67	95

Table 9. Monte Carlo percent relative bias, percent coefficient of variation and percent relative efficiency of several estimators under three sampling designs of size $n = 100$, with a rejection rate equal to 90% and for a log-normal distribution of x.

Model		Basic	*Fuller*	MC	MC-Reg	Cube	Cube-Reg	SRS	SRS-Reg
	RB	-0	-0	-0	-0	-0	-0	0	-0
Linear	CV	9	9	9	9	11	8	30	9
	RE	95	100	89	100	62	104	8	98
	RB	-2	-2	-0	-1	-0	-1	-0	-6
Quadratic	CV	22	22	25	24	24	22	33	22
	RE	99	100	76	89	85	104	44	93
	RB	-14	-14	-2	-7	-0	-6	-2	-29
Exponential	CV	134	134	154	146	143	135	148	125
	RE	100	100	76	85	88	98	83	110
	RB	-3	-3	-0	-1	-0	-1	-0	-8
Bump	CV	28	28	32	29	30	27	41	28
	RE	99	100	77	90	86	104	46	93
	RB	1	1	0	0	0	0	-0	2
Anova	CV	9	9	9	9	8	9	10	9
	RE	100	100	101	97	110	107	83	91
	RB	1	1	0	0	-0	0	0	1
Logistic	CV	11	11	11	11	11	11	14	11
	RE	100	100	102	99	101	100	69	96

and x not linear (e.g., quadratic and exponential), the estimator MC-reg showed some bias, especially for small sample sizes, which can be attributed to the problem of small sample bias. For example, for $n = 50$, MC-reg showed a value of RB of approximately -26% for the exponential relationship and a lognormal distribution for the x-values, see Table 6.

The estimator Cube showed virtually no bias in all the scenarios, as expected. However, it was generally less efficient than Fuller. For example, when the x-values were normal, Cube showed a value of RE of approximately 86% in the case of a linear relationship between y and x, and $n = 25$, see Table 1. Some exceptions occurred in the case of ANOVA and the logistic populations, e.g., see Table 9. The estimator Cube-reg showed virtually no bias regardless of the sample size n when the x-values were normal. In this case, its efficiency was almost identical to that of Fuller with values of RE ranging from 95% to 103%. When the distribution of the x-values was a mixture or log-normal, Cube-reg showed some bias in some scenarios. For example, when the x-variable was log-normal, Cube-reg showed a value of RB oof approximately -38% for the exponential distribution, and for $n = 25$, see Table 3. However, the bias of Cube-reg was significantly smaller than that of Fuller, which exhibited a value of RB close to -71% in the same scenario. The bias of Cube-reg can be attributed to a small sample bias that comes mostly from the fact that the balancing constraints were not exactly satisfied. The same observation can be made when the x-variable was distributed according to a mixture. However, in terms of RE, Fuller was slightly better than Cube-reg. For example, when the x-values were distributed according to a mixture, the values of RE ranged from 90% to 101%, see Tables 2, 5, and 8.

Finally, we discuss the properties of the estimators SRS and SRS-reg obtained under simple random sampling without replacement. As expected, SRS showed virtually no bias in all the scenarios. However, it was generally less efficient than Fuller. On the other hand, SRS-reg was generally better than SRS in terms of RE, although it suffered from small sample bias in some scenarios. For example, when the x-values log-normal, SRS-reg showed a value of RB of approximately -67% for the exponential distribution, and for $n = 25$, see Table 3. On the other hand, Cube-reg showed a value of RB of approximately -38% for the same scenario. This can be explained by the fact that both estimators include the term $(\widehat{\mathbf{B}}^{\pi} - \mathbf{B}^{\pi})^{\top}(\mathbf{t}_{\mathbf{z}} - \widehat{\mathbf{t}}_{\mathbf{z}}^{\pi})$, which is $O_p(Nn^{-3/2})$ for the Cube method and $O_p(Nn^{-1/2})$ for simple random sampling without replacement.

6. Concluding Remarks

In this paper, we examined the properties of several point and estimation procedures. The estimator based on Monte Carlo approximations were generally inefficient. To cope with this problem, it would be interesting to smooth these probabilities through the use of classes formed on the basis on the Monte Carlo approximations. This requires further research.

The properties of the regression estimator advocated by Fuller (2009) depend partly on the rejection rate. For a high rejection (which is typically what would one use in practice to achieve near balance), there are very minor differences between the Horvitz-Thompson type estimator based on the basic inclusion probabilities and the regression estimator of Fuller (2009). Only for low to medium rejections rates, would the use of a regression type estimator improve the efficiency of the estimation procedure significantly.

For the rejective sampling procedure of Fuller (2009), the basic inclusion probabilities p_i are known and fixed prior to sampling. On the other hand, the inclusion probabilities with respect to the rejective sampling design, π_i, are unknown. Another approach consists of fixing the π_i's and determining the basic inclusion probabilities p_i so that, after performing the rejective sampling procedure, the π_i's are exactly or approximately satisfied. In the context of Conditional Poisson sampling, this approach was studied by Dupacova (1979) and Chen, Dempster and Liu (1994). The extension to general rejective sampling procedures is currently under investigation.

Supplementary Materials

The Supplement provides an empirical illustration of both sources of bias in (3.10) (Section 1 of the Supplement), additional results for a rejection rate of 50% (Section 2 of the Supplement) and a proof of (4.3) (Section 3 of the Supplement).

Acknowledgements

The authors are grateful to an associate editor and a reviewer for their comments and suggestionsthat substantially 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

Breidt, F. J. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the Cube method. *J. Statist. Plann. Inference* **141**, 479-487.

Chauvet, G. (2011). On variance estimation for The French master sample. *J. Official Statist.* **27**, 651-668.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm for balanced sampling. *Comput. Statist.* **21**, 53-61.

Chen, X. H., Dempster, A. P. and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81**, 457-469.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.

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

Dupacova, J. (1979). A note on rejective sampling. *Contributions to Statistics* (J. Hájek memorial volume), 71-78, Academia Prague.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika* **93**, 269-278.

Fuller, W. A. (2009). Some design properties of a rejective sampling procedure. *Biometrika* **96**, 933-944.

Hájek, J. (1981). *Sampling from a Finite Population*. Marcel Dekker, New York.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing survey data. *Statist. Sinica* **24**, 375-394.

Kott, P. S. and Liao, D. (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods* **6**, 105-111.

Legg, J. C. and Yu, C. L. (2010). A comparison of sample set restriction procedures. *Surv. Methodol.* **36**, 69-79.

Lesage, E. (2013). Utilisation d'information auxiliaire en théorie des sondages à l'étape de l'échantillonnage et à l'étape de l'estimation. Ph.D. Dissertation, University of Rennes 1.

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

Thompson, M. E. (1997). *Theory of Sample Surveys*. Chapman and Hall.

Thompson, M. E. and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Surv. Methodol.* **34**, 3-10.

Tillé Y. (2011). *Sampling Algorithms*. Springer, New York.

ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France.

E-mail: guillaume.chauvet@ensai.fr

Département de mathématiques et de statistique, Université de Montréal, Québec, H3C 3J7, Canada.

E-mail: david.haziza@umontreal.ca

INSEE, 18 bd Adolphe Pinard 75 014 Paris, France.

E-mail: eric.lesage@insee.fr

(Received August 2013; accepted December 2015)