n° 2010-04

Improved Variance Estimation for Balanced Samples Drawn Via the Cube Method

F. J. BREIDT[1]
G. CHAUVET[2]

[1] Colorado State University, USA.
[2] Ecole Nationale de la Statistique et de l'Analyse de l'Information (ENSAI) CREST, Laboratoire de Statistique d'Enquête, Campus de Ker Lann, 35170 Bruz, France. Email : chauvet@ensai.fr

Improved Variance Estimation for Balanced Samples Drawn Via the Cube Method

July 26, 2010

F. Jay Breidt
Colorado State University

Guillaume Chauvet
Ecole Nationale de la Statistique et de l'Analyse de l'Information
Laboratoire de Statistique d'Enquête

ABSTRACT

The Cube method proposed by Deville and Tillé (2004) enables the selection of balanced samples : that is, samples such that the Horvitz-Thompson estimators of auxiliary variables match the known totals of those variables. As an exact balanced sampling design often does not exist, the Cube method generally proceeds in two steps : a "flight phase" in which exact balance is maintained, and a "landing phase" in which the final sample is selected while respecting the balance conditions as closely as possible. Deville and Tillé (2005) derive a variance approximation for balanced sampling that takes account of the flight phase only, whereas the landing phase can prove to add non-negligible variance. This paper uses a martingale difference representation of the cube method to construct an efficient simulation-based method for calculating approximate second-order inclusion probabilities. The approximation enables nearly unbiased variance estimation, where the bias is primarily due to the limited number of simulations. In a Monte Carlo study, the proposed method has significantly less bias than the standard variance estimator, leading to improved confidence interval coverage.

Keywords : Balanced sampling ; Cube method ; Inclusion probabilities ; Martingale.

1 Introduction

A balanced sampling design has the attractive feature that the Horvitz-Thompson estimators of the totals for auxiliary variables, called balancing variables, exactly match the known totals. Deville and Tillé (2004) introduced the *cube method*, which enables the selection of exact balanced samples if such samples may be found, or approximate balanced samples otherwise. The cube method proceeds in two phases : the *flight phase*, in which balancing constraints are maintained exactly, and the *landing phase*, in which the balancing constraints are successively relaxed until the sample is completed. In the case of balanced sampling with maximum entropy, Deville and Tillé (2005) proposed various variance approximations and associated variance estimators, and compared their performance through a set of simulations.

The variance approximations proposed by Deville and Tillé (2005) do not take into account the whole sampling process, as they ignore the contributions from the landing phase. In fact, the landing phase generates an additional term of variance that may not be negligible in many cases. For example, if the balancing variables have high predictive power for the variable of interest, then the contribution from the landing phase may represent a major part of the variance.

In this paper, we propose a simulation-based approximation of the variance-covariance matrix of the sampling design. The simulation method relies on a martingale difference representation of the cube method. The proposed simulation-based approximation is shown to be particularly efficient as compared to a naive simulation-based approximation, which ignores the martingale difference structure. The approximation enables the computation of a

variance estimator that takes into account the whole sampling process. We show that the resulting variance estimator has significantly less bias than the standard variance estimator from Deville and Tillé (2005), leading to confidence intervals with better coverage rates.

The paper is organized as follows. After defining notation in Section 2, we present the simulation-based variance approximation and the resulting variance estimator in Section 3. The martingale-difference based variance approximation is compared to the naive simulation-based approximation in Section 4.1, and the variance estimator is compared to the standard estimator of Deville and Tillé (2005) in Section 4.2 through a set of simulations. A brief discussion follows in Section 5.

2   Balanced Sampling

Let U denote a finite, labeled population of size N. Let S denote a random sample selected from U by means of a sampling design $p(\cdot)$. Let $\pi_k = \Pr[k \in S]$ denote the inclusion probability of unit k, and $\pi_{kl} = \Pr[k, l \in S]$ denote the second-order inclusion probability. Write $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N)'$ for the vector of inclusion probabilities and $\boldsymbol{I}(S) = (I_1, \ldots, I_k, \ldots, I_N)'$ for the vector of sample membership indicators, where $I_k = 1$ if $k \in S$ and 0 otherwise. The sampling design is assumed to be of fixed size. This implies that $\sum_{k \in U} \pi_k = n = \sum_{k \in U} I_k$, where n denotes the sample size.

The Horvitz-Thompson (1952) estimator,

$$\hat{t}_{\mathbf{z}\pi} = \sum_{k \in U} \frac{\mathbf{z}_k}{\pi_k} I_k, \tag{1}$$

2

estimates without bias the finite population total $t_{\mathbf{z}} = \sum_{k \in U} \mathbf{z}_k$ of the (vector) variable $\mathbf{z}_k$. In the particular case of a scalar variable of interest $\mathbf{z}_k = y_k$, the variance of the Horvitz-Thompson estimator may be obtained by means of the Yates-Grundy (1953) formula :

$$\mathrm{Var}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k,l \in U : k \neq l} \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2, \tag{2}$$

where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

We assume that a vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{qk})'$ of q auxiliary variables is known at the design stage for each unit k in the population. Such auxiliary information allows for possible gains in efficiency through the selection of balanced samples, drawn via the *cube method* (Deville and Tillé, 2004). The sampling design $p(\cdot)$ is said to be *balanced* on variables $\mathbf{x}$ if the equations

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}} \tag{3}$$

hold exactly. That is, the Horvitz-Thompson estimator for the auxiliary vector $\mathbf{x}$ exactly matches the known vector of totals. The variables $\mathbf{x}_k$ are called the balancing variables. For example, the condition of fixed sample size is met if the inclusion probability π_k belongs to the vector $\mathbf{x}_k$ of balancing variables. If the equations (3) are satisfied, the variance of the Horvitz-Thompson estimator is zero for any linear combination of the balancing variables $\mathbf{x}$. As an exact balanced sample generally cannot be found, the cube method enables the selection of approximately balanced samples, proceeding in a flight phase and a landing phase as noted above.

Several implementations of the cube method have been proposed in the lite-

rature. Algorithm 1 given in the appendix covers both the flight phase and the landing phase. The flight phase given in Algorithm 1 corresponds to the general flight phase as proposed by Deville and Tillé (2004) and described by Chauvet and Tillé (2006, Algorithm 1). The variance estimation strategy proposed in this paper may still be applied if the very fast implementation of the flight phase proposed by Chauvet and Tillé (2006, Algorithm 2) is used instead. The landing phase given in Algorithm 1 proceeds by successively relaxing the balancing constraints. A landing phase by means of an enumerative algorithm could alternatively be used (see Tillé, 2006, p. 163), but the variance estimation strategy proposed in §3.1 of this paper would no longer be valid, since the martingale difference representation of the cube method would not be maintained during the whole sampling process. Also, the landing phase given in Algorithm 1 has the advantage of permitting the selection of a balanced sample in a reasonable amount of time, even if the number of balancing variables is large (see Tillé, 2006, p. 164).

Algorithm 1 proceeds in steps $t = 0, 1, \ldots, T$ from $\boldsymbol{\pi}_0(S) = \boldsymbol{\pi}$ to $\boldsymbol{\pi}_T(S) = \boldsymbol{I}(S)$, the final sample. At each step, one or more coordinates of $\boldsymbol{\pi}_t(S)$ are randomly rounded to 0 or 1, and remain there forever. Let $X = [\boldsymbol{x}'_k]_{k \in U}$, the $N \times q$ matrix of auxiliary variables. During the flight phase, the balancing equations remain exactly respected, that is :

$$X' \text{diag}\{\pi_k^{-1}\}_{k \in U} \boldsymbol{\pi}_t(S) = X' \mathbf{1}_N$$

where $\mathbf{1}_N$ denotes the $N \times 1$ vector of ones. When exact balance is no longer possible, the constraints are relaxed successively in the landing phase. It is thus necessary to order the balancing variables with respect to their relative

4

importance, since the last balancing variables are removed first. We assume that this ordering has been done prior to implementation of Algorithm 1.

The cube method belongs to the family of martingale algorithms (Tillé (2006), p. 32). That is, the sequence $\{\boldsymbol{\pi}_t(S)\}_{t=0,\ldots,T}$, where T denotes the stopping time of the algorithm, defines a discrete time martingale.

3 Variance estimation for balanced sampling

The variance of the Horvitz-Thompson estimator is given by formula (2). Consequently, an unbiased variance estimator is given by

$$v\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k,l \in S: k \neq l} \frac{\Delta_{kl}}{\Delta_{kl} + \pi_k \pi_l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \tag{4}$$

if all $\pi_{kl} = \Delta_{kl} + \pi_k \pi_l$ are strictly positive. Second-order inclusion probabilities are, however, usually difficult to compute for a general balanced sampling design. A particular case of balancing variables for which these probabilities may be calculated is when $\mathbf{x}_k = x_k = \pi_k$, that is, when the constraint of fixed sample size is the only balancing constraint. If this balanced sampling design is performed with maximum entropy, Deville (2000) and Matei and Tillé (2004) proposed an algorithm that enables the computation of the exact second-order inclusion probabilities. In general, we resort to simulation for computation of Δ_{kl}.

3.1 Simulation-based variance estimation

We consider two variance estimators, based on two different simulation-based approximations of the design variance-covariance matrix,

$$\Delta = [\Delta_{kl}]_{k,l \in U} = \mathrm{Var}\left(\boldsymbol{I}(S)\right).$$

Since

$$\mathrm{E}\left[\left(\boldsymbol{I}(S) - \boldsymbol{\pi}\right)\left(\boldsymbol{I}(S) - \boldsymbol{\pi}\right)'\right] = \Delta,$$

an obvious, unbiased simulation-based estimator of Δ is given by

$$\Delta_{SIM} = \frac{1}{C} \sum_{c=1}^{C} \left(\boldsymbol{I}(S_c) - \boldsymbol{\pi}\right)\left(\boldsymbol{I}(S_c) - \boldsymbol{\pi}\right)', \tag{5}$$

where $S_1, \ldots, S_c, \ldots, S_C$ are C independent replicates of the sample. These replicates may be selected by Algorithm 1.

The corresponding variance estimator for a given sample S is then obtained by plugging (5) into (4),

$$v_{SIM}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k,l \in S: k \neq l} \frac{\Delta_{SIM,kl}}{\Delta_{SIM,kl} + \pi_k \pi_l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{6}$$

This estimator is no longer exactly unbiased, because the unbiased approximations of the Δ_{kl} enter non-linearly. The estimator also has greater variance than v due to the simulation. Both the bias and the additional variance can be made arbitrarily small for sufficiently large C.

We now propose an alternative simulation-based approximation for Δ that uses the martingale structure of the sampling algorithm. Specifically, note

that

$$I(S) = \pi_T(S) = \pi + \sum_{t=1}^{T} \delta_t(S), \tag{7}$$

where the innovations $\{\delta_t(S)\}_{t=1,\dots,T}$ are given in Algorithm 1. By construction, $\{\delta_t(S)\}$ is a martingale difference (MD) sequence with respect to the sequence of sigma-fields $\mathcal{F}_{t-1} = \sigma(\delta_0(S), \delta_1(S), \dots, \delta_{t-1}(S))$, and so these random vectors are uncorrelated and have mean zero. Hence,

$$
\begin{aligned}
\Delta = \mathrm{Var}(I(S)) &= \mathrm{Var}(\pi_T(S)) \\
&= \sum_{t=1}^{T} \mathrm{Var}(\delta_t(S)) \\
&= \sum_{t=1}^{T} \mathrm{E}\left[\mathrm{Var}(\delta_t(S) \mid \mathcal{F}_{t-1})\right] \\
&= \mathrm{E}\left[\sum_{t=1}^{T} \lambda_{1t}^*(S)\lambda_{2t}^*(S)u_t(S)u_t'(S)\right],
\end{aligned}
$$

the last equality following from Step 3 of Algorithm 1. Consequently the Δ matrix is unbiasedly estimated by

$$\Delta_{MD} = \frac{1}{C} \sum_{c=1}^{C} \sum_{t=1}^{T} \lambda_{1t}^*(S_c)\lambda_{2t}^*(S_c)u_t(S_c)u_t'(S_c), \tag{8}$$

where $S_1, \dots, S_c, \dots, S_C$ are C independent replicates of the sample, selected by Algorithm 1.

The corresponding variance estimator for a given sample S is then obtained by plugging (8) into (4),

$$v_{MD}\left(\hat{t}_{y\pi}\right) = -\frac{1}{2} \sum_{k,l \in S: k \neq l} \frac{\Delta_{MD,kl}}{\Delta_{MD,kl} + \pi_k \pi_l} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \tag{9}$$

A key question is whether the MD-based approximation is any better than

the standard simulation-based approximation, SIM. Note that

$$
\begin{aligned}
(\boldsymbol{I}(S) - \boldsymbol{\pi})(\boldsymbol{I}(S) - \boldsymbol{\pi})' \;=\; & \sum_{r=1}^{T}\sum_{t=1}^{T}\boldsymbol{\delta}_r(S)\boldsymbol{\delta}'_t(S) \\
=\; & \sum_{t=1}^{T}\mathrm{E}\left[\boldsymbol{\delta}_t(S)\boldsymbol{\delta}'_t(S)\,|\,\mathcal{F}_{t-1}\right] \\
& + \sum_{t=1}^{T}\left(\boldsymbol{\delta}_t(S)\boldsymbol{\delta}'_t(S) - \mathrm{E}\left[\boldsymbol{\delta}_t(S)\boldsymbol{\delta}'_t(S)\,|\,\mathcal{F}_{t-1}\right]\right) \\
& + \sum_{r\neq t}\boldsymbol{\delta}_r(S)\boldsymbol{\delta}'_t(S) \\
=\; & \sum_{t=1}^{T}\lambda^*_{1t}(S)\lambda^*_{2t}(S)\boldsymbol{u}_t(S)\boldsymbol{u}'_t(S) + \Lambda_1(S) + \Lambda_2(S),
\end{aligned}
$$

so that

$$
\begin{aligned}
\Delta_{SIM} \;=\; & \Delta_{MD} + \frac{1}{C}\sum_{c=1}^{C}\Lambda_1(S_c) + \frac{1}{C}\sum_{c=1}^{C}\Lambda_2(S_c) \\
=\; & \Delta_{MD} + \Lambda_1 + \Lambda_2.
\end{aligned}
\tag{10}
$$

Clearly $\mathrm{E}\left[\Lambda_1\right] = \mathrm{E}\left[\Lambda_2\right] = 0$. Equation (10) implies, in some sense, that one part of the variability of Δ_{SIM} vanishes in Δ_{MD}, since the two random terms Λ_1 and Λ_2 are omitted. The sign of the covariance between Δ_{MD} and $\Lambda_1 + \Lambda_2$ would be necessary to make this point rigorous; unfortunately we have not found a tractable approach for computation of this covariance. However, our simulation results on a small population suggest that our approximation performs particularly well : see Section 4.1.

3.2 Maximum entropy variance estimation

Another variance approximation is provided by Deville and Tillé (2005). They assume that the balanced sampling is performed with maximum entropy, and that the sampling design is exactly balanced. Then, under an assumption of asymptotic normality of the multivariate Horvitz-Thompson estimator $\hat{t}_{\mathbf{z}\pi}$ under Poisson sampling, they derive the following variance approximation :

$$\text{Var}_{DT}(\hat{t}_{y\pi}) = \frac{N}{N-q} \sum_{k \in U} \pi_k(1-\pi_k) \left(\frac{y_k}{\pi_k} - \frac{y_k^*}{\pi_k} \right)^2, \tag{11}$$

where

$$y_k^* = \mathbf{x}_k' \left(\sum_{l \in U} \pi_l(1-\pi_l)\frac{\mathbf{x}_l\mathbf{x}_l'}{\pi_l^2} \right)^{-1} \sum_{l \in U} \pi_l(1-\pi_l)\frac{\mathbf{x}_l y_l}{\pi_l^2}$$

is a weighted prediction of y_k obtained with the q balancing variables $\mathbf{x}_k$. Other slightly different approximations are proposed in Deville and Tillé (2005), but their simulation results suggest that approximation (11) performs well among variance approximations that may be computed in case of any set of inclusion probabilities. A variance estimator is obtained through a substitution of each total in (11) by its Horvitz-Thompson estimator, using a plug-in principle. The resulting estimator is

$$v_{DT}(\hat{t}_{y\pi}) = \frac{n}{n-q} \sum_{k \in S} (1-\pi_k) \left(\frac{y_k}{\pi_k} - \frac{\tilde{y}_k^*}{\pi_k} \right)^2, \tag{12}$$

where

$$\tilde{y}_k^* = \mathbf{x}_k' \left(\sum_{l \in S} (1-\pi_l)\frac{\mathbf{x}_l\mathbf{x}_l'}{\pi_l^2} \right)^{-1} \sum_{l \in S} (1-\pi_l)\frac{\mathbf{x}_l y_l}{\pi_l^2}.$$

The hypothesis of exact balancing assumed by Deville and Tillé (2005) implies that approximation (11) accounts for the variance due to the flight

phase only. Consequently, the variance estimator given in (12) may lead to serious bias in variance estimation if the variance due to the landing phase is appreciable. With the method we propose, the whole sampling process is taken into account in the variance estimation.

4 Results of the numerical studies

4.1 Comparison of the two simulation-based approximations

We first consider the two simulation-based approximations, Δ_{SIM} and Δ_{MD}, for a small finite population in which exact computation of Δ is possible. The population U is of size $N = 10$. Two balanced designs are considered, one of size $n = 3$ and one of size $n = 5$. For each design, the inclusion probability vector $\boldsymbol{\pi}$ is obtained by drawing $z_1, \ldots, z_N$ as independent and identically distributed Uniform(0,0.2) random variables. Then define $\pi_k = nz_k / \sum_{k \in U} z_k$. We consider the sampling design with $x_{k1} = 1$ and $x_{k2} = \pi_k$ as balancing variables. This is not an exact balanced sampling design.

Since the population is of very small size, the complete support of the design can be enumerated, and the design probabilities can be computed exactly. It turns out that there are 84 possible samples, with probabilities of selection ranging from 3.47×10^{-5} to 3.91×10^{-2}. Further, Δ can be computed exactly.

The two simulation-based approximations Δ_{SIM} and Δ_{MD} given by formulas (5) and (8), respectively, are compared. Following Deville and Tillé (2005), we note that if we approximate Δ by Δ_{approx}, the worst-case variance ap-

proximations are given when the ratio

$$\frac{\boldsymbol{y}'\Delta_{approx}\boldsymbol{y}}{\boldsymbol{y}'\Delta\boldsymbol{y}}$$

is far from one, with extreme values given by

$$\min_{\mathbf{y}} \mathbf{y}'\Delta_{approx}\mathbf{y} \text{ subject to } \mathbf{y}'\Delta\mathbf{y} = 1,$$

and

$$\max_{\mathbf{y}} \mathbf{y}'\Delta_{approx}\mathbf{y} \text{ subject to } \mathbf{y}'\Delta\mathbf{y} = 1.$$

In turn, these optimization problems correspond to choosing the smallest and largest nonnegative general eigenvalues α_{min} and α_{max}, respectively, where the general eigenvalues are the roots of the polynomial $\det\left[\Delta_{approx} - \alpha\,\Delta\right]$. See Deville and Tillé (2005, p. 576) and Harville (1997, pp. 562, 581). A good approximation of the Δ matrix should give general eigenvalues α_{min} and α_{max} close to 1.

These general eigenvalues for Δ_{SIM} and Δ_{MD}, computed with varying numbers of simulations ranging from $C = 50$ to $C = 10,000$, are given in Figure 1 with a logarithmic scale for the eigenvalues. We note that with the proposed MD approximation, α_{min} and α_{max} converge quickly to 1 as the number of simulations grows. That is, the matrix Δ_{MD} converges quickly to Δ. Moreover, the matrix Δ_{MD} systematically outperforms the matrix Δ_{SIM}, regardless of the number of simulations used.
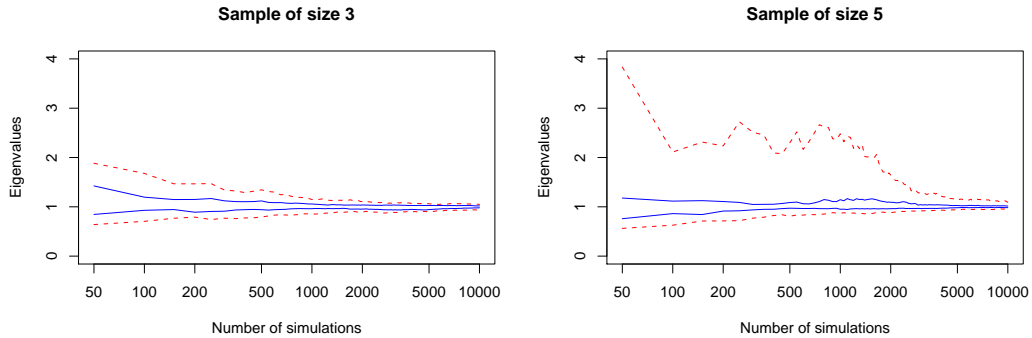
FIGURE 1 – Largest and smallest nonnegative general eigenvalues versus number of simulations, C. The blue, solid curves correspond to the proposed approximation (MD), and the red, dashed curves to the usual simulation-based approximation (SIM).

4.2 Comparison with the Deville-Tillé variance estimator

Some numerical studies adapted from Deville and Tillé (2005) are considered to assess the proposed MD variance estimator. We consider balanced sampling designs in two different populations. In each population, the inclusion probabilities are generated by using a uniform distribution.

Balanced design 1 : The population U_1 is of size $N = 40$. The design considered has unequal probabilities and $n = 15$. For $k \in U_1$, define

$$z_{2k} = k, z_{3k} = k^{-1}, z_{4k} = k^{-2}$$

and let $\bar{z}_i = N^{-1} \sum_{k \in U_1} z_{ik}$, $s_{zi}^2 = (N-1)^{-1} \sum_{k \in U_1} (z_{ik} - \bar{z}_i)^2$ denote the usual empirical mean and variance for $i = 2, 3, 4$. Then $\boldsymbol{x}'_k = (x_{1k}, x_{2k}, x_{3k}, x_{4k})$, where $x_{1k} = \pi_k$ and

$$x_{ik} = \frac{z_{ik} - \bar{z}_i}{s_{zi}}$$

12

for $i = 2, 3, 4$. This is not an exact balanced sampling design.

Balanced design 2 : The population U_2 is of size $N = 30$. The design consi-
dered has unequal probabilities and $n = 10$. Define $x_{k1} = 1$, $x_{k2} = 1$ if
$k \in \{1, \ldots, 15\}$ and 0 otherwise, $x_{k3} = 1$ if $k \in \{11, \ldots, 25\}$ and 0 otherwise,
and $x_{k4} = 1$ if $k \in \{1, \ldots, 5\} \cup \{21, \ldots, 30\}$ and 0 otherwise. This is not an
exact balanced sampling design.

In each population, five variables of interest $y_1, \ldots, y_5$ are generated according
to the linear regression model

$$y_{jk} = \beta_1 + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \sigma_j \epsilon_k, \tag{13}$$

for $j = 1, \ldots, 5$. The ϵ_k's were generating according to a normal distribution
with mean 0 and variance 1. For population U_1, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$. For
population U_2, $\beta_1 = 0$ and $\beta_2 = \beta_3 = \beta_4 = 1$. In each case, the coefficient σ_j
was chosen to give a model R^2 (coefficient of determination) approximately
equal to 0.1 for y_1, 0.2 for y_2, and so on.

Our objective is to estimate the variance of the Horvitz-Thompson estimator
of the totals of the five y-variables of interest. The methods considered in-
cluded the proposed Martingale Difference variance estimator (MD) and the
approximation given by Deville and Tillé (DT). The Δ_{MD} matrix in (8) was
obtained from a separate simulation run of $C = 100,000$ samples. The same
simulation run was used to compute the variance of the Horvitz-Thompson
estimators.

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

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

where $B = 2,000$ independent samples are selected according to the balanced sampling design, and $\hat{\theta}_{(b)}$ gives the value of the estimator for the bth sample. As a measure of variance of an estimator $\hat{\theta}$ we used the Monte Carlo percent relative stability (RS) given by

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

When $\theta = \text{Var}\left(\hat{t}_{y_j \pi}\right), j = 1, \ldots, 5$, we have $\hat{\theta}$ equal to either $v_{MD}(\hat{t}_{y_j \pi})$ or $v_{DT}(\hat{t}_{y_j \pi})$, $j = 1, \ldots, 5$.

We also assess the coverage of confidence intervals computed using the t distribution with $n - q$ degrees of freedom, where q denotes the number of balancing variables. The results are given in Tables 1 and 2 for populations U_1 and U_2, respectively.

In each of the two populations, the relative bias of the DT estimator increases as the R^2 of the model increases, whereas the MD estimator remains approximately unbiased. As a consequence, the DT estimator is also outperformed by the simulation-based variance estimator in terms of confidence interval error rates, especially when R^2 is high. Our interpretation is that R^2 measures the relative importance between the variance due to the flight phase and the variance due to the landing phase in the overall variance. When R^2 is small, the balancing variables have little explanatory power for the variable y. The

| | | Nominal one-tailed error rate | | | | | | % Rel. Bias | % Rel. Stab. |
| | | 2.5 % | | | 5 % | | | | |
Variable	Method	L	U	$L+U$	L	U	$L+U$	RB_{MC}	RS_{MC}
y_1	MD	2.05	2.80	4.85	4.35	5.65	10.00	0.2	33.8
	DT	2.75	2.25	5.00	4.90	5.15	10.05	-0.8	29.3
y_2	MD	1.70	3.20	4.90	3.95	6.65	10.60	0.0	44.3
	DT	3.20	2.70	5.90	6.00	5.80	11.80	-11.5	28.6
y_3	MD	1.35	4.00	5.35	3.25	7.65	10.90	-0.2	57.4
	DT	3.95	3.30	7.25	7.70	6.85	14.55	-22.8	32.2
y_4	MD	1.10	4.70	5.80	2.65	8.80	11.45	-0.4	71.7
	DT	5.70	4.25	9.95	9.25	7.95	17.20	-34.2	39.3
y_5	MD	0.85	5.90	6.75	2.15	10.90	13.05	-0.6	86.5
	DT	7.10	5.20	12.30	11.40	9.75	21.15	-45.7	48.4

TABLE 1 – Monte-Carlo error rates in the upper and lower tails, percent relative biases and relative stabilities of two variance estimators for balanced sampling in population U_1

Variable	Method	Nominal one-tailed error rate						% Rel. Bias RB_{MC}	% Rel. Stab. RS_{MC}
		2.5 %			5 %				
		L	U	$L+U$	L	U	$L+U$		
y_1	MD	1.35	6.45	7.80	3.25	10.55	13.80	-0.2	85.6
	DT	2.40	7.40	9.80	5.20	12.55	17.75	-26.9	60.0
y_2	MD	1.30	6.95	8.25	3.20	10.80	14.00	0.0	86.0
	DT	2.50	7.55	10.05	5.05	12.70	17.75	-28.8	59.6
y_3	MD	1.80	7.20	9.00	3.65	11.15	14.80	0.1	86.3
	DT	2.75	8.05	10.80	5.25	12.70	17.95	-31.2	59.3
y_4	MD	1.90	7.05	8.95	4.20	11.25	15.45	0.1	86.8
	DT	2.85	8.15	11.00	5.65	13.25	18.90	-34.2	59.2
y_5	MD	2.40	7.50	9.90	4.30	11.30	15.60	0.2	87.5
	DT	3.05	8.85	11.90	6.25	13.45	19.70	-38.0	59.3

TABLE 2 – Monte-Carlo error rates in the upper and lower tails, percent relative biases and relative stabilities of two variance estimators for balanced sampling in population U_2

variance due to the flight phase is then high (close to the variance obtained in case of a simple fixed-size sampling strategy) and dominates the overall variance. When R^2 becomes larger, the balancing variables have more and more explanatory power for the variable of interest, and the importance of the variance due to the flight phase in the overall variance decreases. The extreme case is when $R^2 = 1$. The variable of interest y is then perfectly explained by the balancing variables. In this case, the variance due to the flight phase is 0. Because only the flight phase is taken into account in the DT variance estimator, it is therefore more and more negatively biased as R^2 increases. This effect is seen in both U_1 and U_2.

The MD variance estimator has considerable variability, in terms of percent relative stability, partly due to the small sample size considered in these examples. The DT variance estimator has low variability, but its extreme bias makes this point irrelevant in all cases except y_1 for population U_1, where DT has low bias. In this case, the relative stability of DT (29.3%) is comparable to that of MD (33.8%).

Note that the variance of the Horvitz-Thompson estimator may be decomposed as follows :

$$
\begin{aligned}
\mathrm{Var}\left(\hat{t}_{y\pi}\right) &= \mathrm{Var}\left(\mathrm{E}\left[\hat{t}_{y\pi} \,|\, \boldsymbol{\pi}_{T(F)}(S)\right]\right) + \mathrm{E}\left[\mathrm{Var}\left(\hat{t}_{y\pi} \,|\, \boldsymbol{\pi}_{T(F)}(S)\right)\right] \\
&= \qquad\qquad V_F \qquad\qquad + \qquad\qquad V_L,
\end{aligned}
$$

where $\boldsymbol{\pi}_{T(F)}(S)$ stands for the sequence $\boldsymbol{\pi}_t(S)$ at the end of the flight phase, $t = T(F)$. The term V_F stands for the variance due to the flight phase, whereas the term V_L stands for the variance due to the landing phase. The term V_L is ignored in the DT estimator, which results in a negatively-biased estimator. However, $\mathrm{Var}\left(\hat{t}_{y\pi} \,|\, \boldsymbol{\pi}_{T(F)}(S)\right)$ is highly unstable, since the units

remaining at the end of the flight phase as well as their number are random. As a consequence, the MD estimator which helps in tracking the term V_L is also highly variable. This variability is needed to achieve approximately correct confidence interval coverage.

5   Conclusion

In this paper, a new variance estimator for balanced sampling, using a martingale difference representation of the cube method, is proposed. This estimator is shown to perform well as compared to a naive simulation-based variance estimator that does not use the martingale structure. Unlike the Deville and Tillé (2005) variance estimator, which ignores sampling variation due to the landing phase, the proposed estimator takes into account the complete sampling process and leads to essentially unbiased variance estimators, and corresponding confidence intervals with proper coverage. Numerical results support our findings.

References

Chauvet, G., and Tillé, Y. (2006). *A Fast Algorithm of Balanced Sampling.* Computational Statistics, 21, pages 53-61.

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

Deville, J-C. (2000). *Note sur l'algorithme de Chen, Dempster et Liu.* Technical Report, Ensai, France.

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

Deville, J-C., and Tillé, Y. (2005). *Variance approximation under balanced sampling.* Journal of Statistical Planning and Inference, 128, pages 569-591.

Harville, D.A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer, New-York.

Horvitz, D.G., and Thompson, D.J. (1952). *A generalization of sampling without replacement from a finite universe.* Journal of the American Statistical Association, 47, pages 663-685.

Matei, A., and Tillé, Y. (2005). *Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size.* Journal of Official Statistics, 21, pages 543-570.

Yates, F., and Grundy, P.M. (1953). *Selection without replacement from within strata with probability proportional to size.* Journal of the Royal Statistical Society B, 15, pages 253-261.

Tillé, Y. (2006). *Sampling Algorithms.* Springer Series in Statistics, New-

York.

Appendix : Algorithm 1 for implementation of the cube method

Define the balancing matrix $A = (\mathbf{a}_1, \ldots, \mathbf{a}_k, \ldots, \mathbf{a}_N)$, where $\mathbf{a}_k = \mathbf{x}_k/\pi_k$. First initialize with $\boldsymbol{\pi}_0(S) = \boldsymbol{\pi}$ and $A(0) = A$. Next, at time $t = 0, \ldots, T$, repeat the three following steps.

Step 1 :

Denote $\boldsymbol{\pi}_t(S) = [\pi_{t,1}(S), \ldots, \pi_{t,k}(S), \ldots, \pi_{t,N}(S)]'$. Let $E(t) = F(t) \cap \mathrm{Ker} A(t)$, where

$$F(t) = \{\mathbf{v} \in \mathbb{R}^n : v_k = 0 \text{ if } \pi_{t,k}(S) \text{ is an integer.}\}$$

– If $E(t) \neq \{0\}$, then generate any vector $\mathbf{u}_t(S) \neq \mathbf{0}$ in $E(t)$, random or not. Put $A(t+1) = A(t)$.

– If $E(t) = \{0\}$, let m_t denote the largest integer such that $F(t) \cap \mathrm{Ker} A_{m_t}(t) \neq \{0\}$, where $A_{m_t}(t)$ denotes the matrix given by the first m_t rows of $A(t)$. Generate any vector $\mathbf{u}_t(S) \neq 0$ in $F(t) \cap \mathrm{Ker} A_{m_t}(t)$, random or not. Put $A(t+1) = A_{m_t}(t)$.

Step 2 :

Compute the scalars $\lambda_{1t}^*(S)$ and $\lambda_{2t}^*(S)$, which are the largest values of λ_{1t}

and λ_{2t} such that

$$0 \leq \boldsymbol{\pi}_t(S) + \lambda_{1t}\mathbf{u}_t(S) \leq 1, \ 0 \leq \boldsymbol{\pi}_t(S) - \lambda_{2t}\mathbf{u}_t(S) \leq 1,$$

where the inequalities are interpreted element-wise. Note that $\lambda_{1t}^*(S) > 0$ and $\lambda_{2t}^*(S) > 0$.

Step 3 :

Select $\boldsymbol{\pi}_{t+1}(S) = \boldsymbol{\pi}_t(S) + \boldsymbol{\delta}_t(S)$, where

$$\boldsymbol{\delta}_t(S) = \begin{cases} \lambda_{1t}^*(S)\mathbf{u}_t(S) & \text{with probability} \quad q(t) \\ -\lambda_{2t}^*(S)\mathbf{u}_t(S) & \text{with probability} \quad 1 - q(t) \end{cases}$$

and $q(t) = \lambda_{2t}^*(S)/(\lambda_{1t}^*(S) + \lambda_{2t}^*(S))$.

The procedure ends at step T, when $\boldsymbol{\pi}_T(S)$ has only integer (0–1) components. It follows from Proposition 1 of Deville and Tillé (2004, p. 897) that the number of steps T is no greater than the size N of the population.