Comput Stat (2014) 29:1345–1363
DOI 10.1007/s00180-014-0495-0

ORIGINAL PAPER

A new resampling method for sampling designs without replacement: the doubled half bootstrap

Erika Antal · Yves Tillé

Received: 16 January 2013 / Accepted: 15 April 2014 / Published online: 8 May 2014

Abstract A new and very fast method of bootstrap for sampling without replacement from a finite population is proposed. This method can be used to estimate the variance in sampling with unequal inclusion probabilities and does not require artificial populations or utilization of bootstrap weights. The bootstrap samples are directly selected from the original sample. The bootstrap procedure contains two steps: in the first step, units are selected once with Poisson sampling using the same inclusion probabilities as the original design. In the second step, amongst the non-selected units, half of the units are randomly selected twice. This procedure enables us to efficiently estimate the variance. A set of simulations show the advantages of this new resampling method.

Keywords Poisson sampling · Simple random sampling · Unequal probability sampling · Variance estimation

1 Introduction

Resampling methods are frequently used to draw inference in survey statistics. The main difficulty, however, is that the variance of an estimator depends on the sampling design. Bootstrap methods must thus be adapted for each sampling design. Moreover, the variance of the Horvitz–Thompson estimator can have a very different form from

E. Antal
Swiss Centre of Expertise in the Social Science,
Quartier UNIL-Mouline, Bâtiment Géopolis, 1015 Lausanne, Switzerland
e-mail: erika.antal@fors.unil.ch

E. Antal · Y. Tillé (✉)
Faculty of Economics and Business, Institute of Statistics, University of Neuchâtel,
Rue de la Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland
e-mail: yves.tille@unine.ch

the variance estimator. The original bootstrap method, developed by Efron (1979) is not directly applicable in sampling from a finite population because the units of the sample are not independent and identically distributed when the sample is selected without replacement. Gross (1980), Booth et al. (1994) and Chao and Lo (1985) proposed a method based on the construction of pseudo-populations from the sample (see also the generalizations of Chauvet 2007). Another important family of methods is the rescaled bootstrap (Rao and Wu 1988) which consists of modifying the values of the interest variable to reconstruct an unbiased variance estimator for statistics that are linear functions of the observations. The Rao and Wu method can however be also presented as a weighting system (Rao et al. 1992) that is applied on the vector of the variable of interest. Other methods were also proposed by Mac Carthy and Snowden (1985), Kuk (1989), Rao et al. (1992), Shao and Tu (1995), Sitter (1992a), Sitter (1992b), Holmberg (1998).

The main idea of this paper is similar to the general weighted bootstrap (Mason and Newton 1992; Bertail and Combris 1997), which has also been used in the paper of Lahiri (2003). Beaumont and Patak (2012) propose a bootstrap method that directly reconstructs the variance for linear cases. Antal and Tillé (2011a) propose another method that uses non-integer weights and that is based on mixture of discrete multivariate distributions. In this paper, we propose a new methodology to select a bootstrap sample for sampling without replacement from a finite population. This method is an extension of the half-sample bootstrap proposed by Saigo et al. (2001) to which we add a correction for finite population in order to correctly estimate the variance in unequal probability sampling. This method enables one to quickly implement and directly reconstruct the appropriate variance without need of reweighting the statistical units. Indeed, each unit is duplicated an integer number of times.

The paper is organized as follows: in Sect. 2, the notation for a sampling design, the estimator of the total and its variance estimator are defined. Section 3 is devoted to the conditions needed to obtain unbiased bootstrap estimates of the variances. In Sect. 4, a new method is proposed for Poisson sampling. Next, the doubled half sampling is defined in Sect. 5. This tool is used to define a new bootstrap method for simple random sampling in Sect. 6 and for unequal probability sampling in Sect. 7. Simulations are presented in Sect. 8 and the interest of this new method is discussed in Sect. 9.

2 Sampling design, total and variance

Let $p(.)$ be a sampling design on a population $U = \{1, \ldots, N\}$ of size N such that

$$p(s) \geq 0, \text{ for all } s \subset U, \text{ and } \sum_{s \subset U} p(s) = 1.$$

Let S be the random sample such that $\Pr(S = s) = p(s)$. The sample size n of S can be random or not. Define also the inclusion probabilities $\pi_k = \Pr(k \in S)$ for $k \in U$, and the joint inclusion probabilities $\pi_{k\ell} = \Pr(k \text{ and } \ell \in S)$ for $k, \ell \in U$. Moreover, define $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$ for $k, \ell \in U$, and $\breve{\Delta}_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$. When $k = \ell$, we obtain $\Delta_{kk} = \pi_k(1 - \pi_k)$, $k \in U$ and $\breve{\Delta}_{kk} = 1 - \pi_k$.

If all the inclusion probabilities are positive, then the total $Y = \sum_{k\in U} y_k$ of the values $y_1, \ldots, y_k, \ldots, y_N$ taken by the interest variable y can be unbiasedly estimated by the Horvitz–Thompson estimator (HT) (Horvitz and Thompson 1952)

$$\widehat{Y}_\pi = \sum_{k\in S} \frac{y_k}{\pi_k}, \tag{1}$$

whose variance is given by

$$\mathrm{var}(\widehat{Y}_\pi) = \sum_{k\in U}\sum_{\ell\in U} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Delta_{k\ell}$$

and can be unbiasedly estimated by the Horvitz–Thompson (HT) variance estimator

$$\widehat{\mathrm{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \check{\Delta}_{k\ell}.$$

This variance estimator is however very unstable because

$$\sum_{k\in S} \check{\Delta}_{k\ell} \tag{2}$$

is generally different from zero even when $\sum_{k\in U} \Delta_{k\ell} = 0$. In the particular case where the sample size is fixed and $y_k = \pi_k, k \in U$, the HT-estimator of the total is equal to the sample size and is thus not random. If (2) is different from zero, the HT-variance estimator is generally not zero. Indeed when $y_k = \pi_k$,

$$\widehat{\mathrm{var}}_{HT}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S} \check{\Delta}_{k\ell}.$$

Thus when (2) is different from zero $\widehat{\mathrm{var}}_{HT}(\widehat{Y}_\pi)$ can be random even when $\mathrm{var}(\widehat{Y}_\pi)$ is null.

When the sample size is fixed, the Sen–Yates–Grundy (SYG) estimator is also unbiased (Sen 1953; Yates and Grundy 1953):

$$\widehat{\mathrm{var}}_{SYG}(\widehat{Y}_\pi) = \sum_{k\in S}\sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} D_{k\ell},$$

where

$$D_{k\ell} = \begin{cases} -\sum\limits_{\substack{j\in S \\ j\neq k}} \frac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \frac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases} \tag{3}$$

Several other variance estimators exist. They all have the same form as the SYG-estimator with different values for $D_{k\ell}$. Matei and Tillé (2005) discussed the merits of a family of estimators based on another value of $D_{k\ell}$ given by:

$$\widetilde{D}_{k\ell} = \begin{cases} c_k - \dfrac{c_k^2}{\sum_{j\in S} c_j} & \text{if } k = \ell \\ -\dfrac{c_k c_\ell}{\sum_{j\in S} c_j} & \text{if } k \neq \ell. \end{cases}$$

Diverse values have been proposed for the c_k. Matei and Tillé (2005) ran a set of simulations that shows that the choice proposed by Hájek (1981):

$$c_k = \frac{n}{n-1}(1-\pi_k). \tag{4}$$

produces a very efficient and slightly biased estimator. We refer to this estimator as the H-estimator of the variance.

3 Bootstrap

A bootstrap sample is a sample with replacement that is not necessarily a simple random sample and that is selected from S. Let S_k^* be the number of times unit k is repeated in the bootstrap sample. The HT estimator of the total for a single bootstrap sample is given by

$$\widehat{Y}^* = \sum_{k\in S} \frac{y_k}{\pi_k} S_k^*.$$

Let $\Pr^*(.) = \Pr(.|S)$, $\mathrm{E}^*(.) = \mathrm{E}(.|S)$, $\mathrm{var}^*(.) = \mathrm{var}(.|S)$ and $\mathrm{cov}^*(.,.) = \mathrm{cov}(.,.|S)$ respectively denote the probability, the expectation, the variance and the covariance operators in the bootstrap sample conditional on the original sample. This gives us

$$\mathrm{E}^*(\widehat{Y}^*) = \sum_{k\in S} \frac{y_k}{\pi_k} \mathrm{E}(S_k^*),$$

and

$$\mathrm{var}^*(\widehat{Y}^*) = \sum_{k\in S}\sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \mathrm{cov}^*(S_k^*, S_\ell^*).$$

A necessary and sufficient condition for the expected value $\mathrm{E}^*(\widehat{Y}^*)$ to equal the HT estimator of the total is

$$\mathrm{E}^*(S_k^*) = 1, k \in S. \tag{5}$$

Moreover, in order to have an unbiased bootstrap estimate of the variance of the HT total estimator, we can define a bootstrap method such that the variance of the

bootstrap estimators of the total is equal to the HT variance estimator given in (1). In order to satisfy this equality, the necessary and sufficient condition is composed of two parts. The first condition is that

$$\text{var}^*(S_k^*) = \check{\Delta}_{kk} = 1 - \pi_k, k \in S. \tag{6}$$

The second condition is that

$$\text{cov}^*(S_k^*, S_\ell^*) = \check{\Delta}_{k\ell}, k \neq \ell \in S. \tag{7}$$

The condition on the covariances is however difficult to meet when the sample is selected with fixed sample size and unequal inclusion probabilities. In this particular case, it is difficult to exactly satisfy more than conditions (5) and (6). Condition (7) can however be approximately satisfied.

When the sample size is fixed, another way of constructing an unbiased estimator of the variance is to equalize the variance of the bootstrap estimator of the total with the SYG-estimators of the variance. The conditions become

$$\text{E}^*(S_k^*) = 1, k \in S, \tag{8}$$
$$\text{var}^*(S_k^*) = D_{kk}, k \in S, \tag{9}$$

and

$$\text{cov}^*(S_k^*, S_\ell^*) = D_{k\ell}, k \neq \ell \in S. \tag{10}$$

But again, conditions (10) on the covariances are difficult to meet when the sample is selected with unequal inclusion probabilities, but it could be approximately satisfied. For unequal probability sampling with fixed sample size, there exist two bootstrap strategies that may be used to approximate either the HT or the SYG-estimator.

The bootstrap estimator of the variance $v_{boot}(\widehat{Y}^*)$ is computed by generating a set of bootstrap samples and by computing the variance of the outcomes of $\widehat{Y}^*$. Moreover, if a bootstrap method provides an approximately unbiased estimator for the variance of totals, it will also provide approximately unbiased variance estimators for smooth functions of totals.

There exists a lot of distributions that satisfy conditions (5) and (6) or (8) and (9). However, if the original sample is selected with unequal inclusion probabilities and if we impose that S_k^* is integer and that the sum of the S_k^* is not random, the problem is really complex. This is a problem of sampling with replacement and with unequal probabilities where the expectation and the variance are fixed. For this reason let us begin with a simple case: Poisson sampling design.

4 Bootstrap for Poisson design

Suppose the original sample is obtained by a so-called Poisson design, where the observation k is included with probability π_k and the decision is made for each observation independently. The name reflects the fact that if all π_k are small, then the sample size

n has approximately a Poisson distribution with mean $\sum_{k=1}^{N} \pi_k$. In a Poisson design with inclusion probabilities π_k,

$$p(s) = \prod_{k=1}^{N} \left[\pi_k^{\mathbb{1}(k\in s)} (1-\pi_k)^{\mathbb{1}(k\notin s)} \right], \text{ for all } s \subset U,$$

where $\mathbb{1}(A)$ is equal to 1 if A is true and 0 otherwise. The inclusion probability is $\Pr(k \in S) = \pi_k$. Moreover, $\pi_{k\ell} = \pi_k \pi_\ell$ when $k \neq \ell \in U$ and $\pi_{kk} = \pi_k$. Thus $\Delta_{k\ell} = 0$, when $k \neq \ell \in U$ and $\Delta_{kk} = \pi_k(1-\pi_k)$. We thus have, $\check{\Delta}_{k\ell} = 0$, when $k \neq \ell \in U$ and $\check{\Delta}_{kk} = 1 - \pi_k$. With Poisson sampling design the sample size n is random thus the estimator of variance is calculated by $\widehat{\text{var}}_{HT}(\widehat{Y}_\pi)$.

Patak and Beaumont (2009) propose a bootstrap method for Poisson design that use several different distribution as normal or lognormal random variables with expectation equal to 1 and variances equal to $1 - \pi_k$. Unfortunately, this method requires the use of non-integer weights. Instead we recommend the use of a discrete random variable for S_k^*.

Antal and Tillé (2011a) propose a simple bootstrap method that uses n independent Bernoulli random variables X_k with parameter π_k and n independent Poisson random variables Z_k with parameter $\lambda = 1$. For this method, the bootstrap sample is given by

$$S_k^* = X_k + (1 - X_k) Z_k, k \in S.$$

Thus, the probability mass function of S_k^* is given by:

$$\Pr^*(S_k^* = r) = \pi_k \mathbb{1}[r = 1] + \frac{(1-\pi_k)}{e \cdot r!}, r = 0, 1, 2, \ldots$$

where $e \approx 2.71$ is the Euler constant. The bootstrap variable S_k^* satisfies conditions (5), (6), and (7).

We propose another method based on n independent Bernoulli random variables $X_k, k \in S$ with parameter π_k and n independent Bernoulli random variables Y_k with parameter $1/2$. Define the bootstrap sample by

$$S_k^* = X_k + 2(1 - X_k) Y_k, k \in S.$$

The probability distribution of S_k^* is thus

$$S_k^* = \begin{cases} 0 & \text{with a probability } (1-\pi_k)/2 \\ 1 & \text{with a probability } \pi_k \\ 2 & \text{with a probability } (1-\pi_k)/2. \end{cases}$$

Again, the bootstrap variable S_k^* meets conditions (5), (6), and (7). Here, the bootstrap sample does not contain the same unit more than twice.

5 One-one design and doubled half sampling

When the original sample has a fixed sample size, Antal and Tillé (2011a) propose one-one designs, a tool in order to estimate the variance of an estimator via bootstrap method. Members of this family are discrete probability distributions with

$$\begin{aligned} \mathrm{E}^*(S_k^*) &= 1, \\ \mathrm{var}^*(S_k^*) &= 1. \end{aligned}$$

The name comes from these conditions on their expectation and variance. Each bootstrap sample has the same fixed sample size

$$\sum_{k \in S} S_k^* = n.$$

The covariance between S_k^* and S_ℓ^* is given by

$$\mathrm{cov}^*(S_k^*, S_\ell^*) = -\frac{1}{n-1}, k \neq \ell \in S.$$

Antal and Tillé (2011a) showed that such a sampling design can be obtained by using a mixture between two samples selected by simple random sampling with replacement and by simple random sampling with over-replacement (Antal and Tillé 2011b). One-one designs can next be mixed with other sampling designs in order to reproduce an unbiased estimator of variance for most of the sampling methods with fixed sample size.

We propose another method for selecting a one-one design that we call "doubled half sampling". In the next section we use this in a new bootstrap procedure. If the size n of the initial sample is *even*, then a sample from S of size $n/2$ is selected with simple random sampling without replacement. Next, each selected unit is taken twice. In this case, we obtain

$$\begin{aligned} \mathrm{E}^*(S_k^*) &= 2 \times \frac{1}{2} = 1, \\ \mathrm{var}^*(S_k^*) &= 4 \times \frac{1}{2}\left(1 - \frac{1}{2}\right) = 1, \end{aligned}$$

and

$$\mathrm{cov}^*(S_k^*, S_\ell^*) = 4 \times \frac{1}{2}\left(1 - \frac{1}{2}\right)\frac{-1}{n-1} = \frac{-1}{n-1}.$$

If n is odd, then we can have the same property by means of the following slightly modified procedure:

- Select $(n-1)/2$ units from S and take them twice in the bootstrap sample.

- With a probability $1/4$, select a unit with equal probabilities from the set of units selected twice. This unit is selected three times.
- Otherwise, with a probability $3/4$, select a unit with equal probabilities among the units that are not selected twice. This unit is selected only once.

This procedure gives the following distribution for S_k^*:

$$\Pr^*(S_k^* = j) = \begin{cases} \frac{n+1}{2n} \times \left(1 - \frac{3}{4} \times \frac{2}{n+1}\right) = \frac{2n-1}{4n} & \text{if } j = 0 \\ \frac{n+1}{2n} \times \frac{3}{4} \times \frac{2}{n+1} = \frac{3}{4n} & \text{if } j = 1 \\ \frac{n-1}{2n} \times \left(1 - \frac{1}{4} \times \frac{2}{n-1}\right) = \frac{2n-3}{4n} & \text{if } j = 2 \\ \frac{n-1}{2n} \times \frac{1}{4} \times \frac{2}{n-1} = \frac{1}{4n} & \text{if } j = 3. \end{cases}$$

After some algebra, it can be shown that this design is one-one. A one-one design can thus be selected for any sample size except when $n = 1$.

6 Bootstrap for simple random sampling without replacement

In this section we propose a bootstrap method for use when original samples are selected by simple random sampling without replacement, where

$$p(s) = \begin{cases} \frac{n!(N-n)!}{N!} & \text{for all the samples } s \text{ of size } n \\ 0 & \text{otherwise.} \end{cases}$$

The inclusion probability is $\pi_k = n/N$. Moreover,

$$\pi_{k\ell} = \frac{n(n-1)}{N(N-1)}$$

when $k \neq \ell \in U$ and $\pi_{kk} = n/N$. Thus,

$$\Delta_{k\ell} = -\frac{n(N-n)}{N^2(N-1)},$$

when $k \neq \ell \in U$ and

$$\Delta_{kk} = \frac{n(N-n)}{N^2}.$$

We thus have,

$$\check{\Delta}_{k\ell} = -\frac{N-n}{N(n-1)},$$

when $k \neq \ell \in U$ and $\check{\Delta}_{kk} = 1 - n/N$. Note also that in this case, the HT-estimator and the SYG-estimator of the variance are equal, i.e. $\check{\Delta}_{kk} = D_{kk} = 1 - n/N$ for all $k \in U$.

We propose using bootstrap samples using the following two-stage procedure. Let S_k^* denotes the number of times unit k is selected in the bootstrap sample.

- Select units from S by using independent Bernoulli random variables $X_k, k \in S$, with probabilities $\pi_k = n/N$. Let $m = \sum_{k \in S} X_k$. Thus $\mathrm{E}(m) = n^2/N$. For now we set $S_k^* = X_k$, including each selected unit in the bootstrap sample once, though this may be adjusted later.
- – If the number of non-selected units is greater than or equal to 2 ($n - m \geq 2$), then select a doubled half sampling design amongst the units $k \in S$ such that $X_k = 0$.
  – If there is exactly one (say unit ℓ) non-selected unit ($n - m = 1$), select that unit with distribution

$$S_\ell^* = \begin{cases} 0 & \text{with probability } 1/4 \\ 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4. \end{cases}$$

  Next, randomly select one of the units such that $X_k = 1$ (say z) with equal probability and select it $S_z^* = 2 - S_\ell^*$ times.
  – If the number of units such that $X_k = 0$ is null ($n - m = 0$), then the bootstrap sample S_k^* is the same as the original sample.

Note that

$$\Pr^*(S_k^* = j | m = r \text{ and } n - r \text{ is even}) = \begin{cases} (1 - r/n)/2 & \text{if } j = 0 \\ r/n & \text{if } j = 1 \\ (1 - r/n)/2 & \text{if } j = 2, \end{cases}$$

$$\Pr^*(S_k^* = j | m = r, n - r \text{ is odd, and } r < n - 1) = \begin{cases} \left(1 - \frac{r}{n}\right) \frac{2n-1}{4n} & \text{if } j = 0 \\ \frac{r}{n} + \left(1 - \frac{r}{n}\right) \frac{3}{4n} & \text{if } j = 1 \\ \left(1 - \frac{r}{n}\right) \frac{2n-3}{4n} & \text{if } j = 2 \\ \left(1 - \frac{r}{n}\right) \frac{1}{4n} & \text{if } j = 3, \end{cases}$$

and

$$\Pr^*(S_k^* = j | m = n - 1) = \begin{cases} 1/(2n) & \text{if } j = 0 \\ 1 - 1/n & \text{if } j = 1 \\ 1/(2n) & \text{if } j = 2. \end{cases}$$

It can be checked that $\mathrm{E}(S_k^*|m) = 1$ and $\mathrm{var}(S_k^*|m) = 1 - m/n$, for all three cases. We obtain $\mathrm{E}(S_k^*) = \mathrm{E}[\mathrm{E}(S_k^*|m)] = 1$ and

$$\begin{aligned} \mathrm{var}(S_k^*) &= \mathrm{E}\ \mathrm{var}(S_k^*|m) \\ &\quad + \mathrm{var}\ \mathrm{E}(S_k^*|m) = 1 - \mathrm{E}(m)/n = 1 - \mathrm{E}(m)/n = 1 - (n^2/N)n = 1 - n/N. \end{aligned}$$

We thus have $\mathrm{E}^*(S_k^*) = 1$, $\mathrm{var}^*(S_k^*) = \breve{\Delta}_{kk}, k \in S$. The values $\mathrm{cov}^*(S_k^*, S_\ell^*) = \breve{\Delta}_{k\ell}$ do not depend on k and ℓ because units have all been treated symmetrically in the sense that there is no particular treatment for a particular unit, which implies that all the covariances $\mathrm{cov}^*(S_k^*, S_\ell^*)$ are equal for all the couples k, ℓ such that $k \neq \ell$.

This procedure differs from the method proposed in Antal and Tillé (2011a). Indeed, in the first stage a Bernoulli design is used whereas in Antal and Tillé (2011a) almost fixed sample size is used. In the second stage, a doubled half sampling whereas in Antal and Tillé (2011a) a complex mixture of distribution is applied. The new method is thus less complex.

7 Bootstrap for unequal probability sampling without replacement

In this section we propose a bootstrap strategy for estimating the variance of an estimator when the original sample is selected by means of an unequal probability sampling design. There is a large set of methods of sampling with unequal inclusion probabilities and fixed sample size (see among others Brewer and Hanif 1983; Tillé 2006). Each method has a particular matrix of $\breve{\Delta}_{k\ell}$ but the diagonal of this matrix is always $\breve{\Delta}_{kk} = 1 - \pi_k, k \in U$. However the sampling designs with large entropy have very similar joint inclusion probabilities (see Brewer and Donadio 2003; Matei and Tillé 2005; Henderson 2006; Preston and Henderson 2007).

Consider the procedure used to compute the inclusion probabilities from a vector of positive values x_k. First, compute the quantities

$$\frac{n x_k}{\sum_{\ell \in U} x_\ell}, \tag{11}$$

$k = 1, \ldots, N$. For units where the quantities are larger than 1, set $\pi_k = 1$. Next, the quantities are recalculated using (11) restricted to the remaining units. This procedure is repeated until each π_k is in $(0, 1]$. Some π_k are 1 and others are proportional to x_k. Let $\pi_k = H_k(x_1, \ldots, x_N; n)$ denote the function that allows us to construct these inclusion probabilities from a vector of positive values $(x_1, \ldots, x_N)$. Function $H(.;.)$ will be used in the new bootstrap we propose for unequal probability sampling without replacement.

A bootstrap method for unequal probability sampling without replacement should satisfy

$$\mathrm{E}^*(S_k^*) = 1,$$

and must have a fixed sample size i.e.

$$\sum_{k \in S} S_k^* = n.$$

Moreover $\mathrm{var}^*(S_k^*)$ should be equal either to the diagonal of matrix $(\breve{\Delta}_{k\ell})$ used for the HT-estimator or to the diagonal of matrix $(D_{k\ell})$ used for the SYG-estimator, i.e.

$$\text{var}^*(S_k^*) = \check{\Delta}_{k\ell} = 1 - \pi_k, k \in S,$$

or, by posing $\phi_k = 1 - D_{kk}$,

$$\text{var}^*(S_k^*) = D_{kk} = 1 - \phi_k, k \in S.$$

With unequal inclusion probabilities, it is difficult to construct a bootstrap method that meets the properties of the covariances given by

$$\text{cov}^*(S_k^*, S_\ell^*) = \check{\Delta}_{k\ell}, \quad \text{or} \quad \text{cov}^*(S_k^*, S_\ell^*) = D_{k\ell}, k \neq \ell \in S. \tag{12}$$

Nevertheless, since the bootstrap sample has a fixed sample size, the relation

$$\sum_{k \in S} \text{cov}^*(S_k^*, S_\ell^*) = 0 \tag{13}$$

is the same as for $D_{k\ell}$. Indeed,

$$\sum_{k \in S} D_{k\ell} = 0, \tag{14}$$

which ensures that (12) will be approximately satisfied when the sampling design has large entropy. Let S_k^* denotes the number of times unit k is selected in the bootstrap sample.

- Select units from S by using a Poisson random variable $X_k, k \in S$, with the same inclusion probabilities as the original design π_k. The selected units are taken once in the bootstrap sample S_k^*. For now we set $S_k^* = X_k$, including each selected unit in the bootstrap sample once, though this may be adjusted later. Let $m = \sum_{k \in S} X_k$. Thus $\text{E}(m) = \sum_{k \in S} \pi_k$.
- – If the number of non-selected units is greater than or equal to 2 ($n - m \geq 2$), then select a doubled half sampling design amongst the units such that $X_k = 0$.
  – If there is exactly one (say unit ℓ) non-selected unit ($n - m = 1$), the S_k^* are completely redefined.
    - With a probability 1/2, select the same bootstrap sample as the original sample ($S_k^* = 1, k \in S$).
    - Otherwise, with a probability 1/2, compute

$$\pi_{k|n-1} = \text{E}(X_k | m = n - 1) = 1 - \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in S} \frac{1-\pi_\ell}{\pi_\ell}}$$

(see the Appendix for the proof). With a sampling method with unequal inclusion probabilities with fixed sample size, select $n - 2$ units from S with probability

$$\psi_k = 1 - H_k(1 - \pi_{k|n-1}, k \in S; 2)$$

and take them once in the bootstrap sample. Select a sample with a doubled half sampling from the two units that are not selected. Note that $\psi_k = 2\pi_{k|n-1} - 1$ except when one of the $\pi_{k|n-1}$ is less than $1/2$.

Note that the procedure proposed in Sect. 6 for simple random sampling is a particular case of the procedure for unequal probability sampling when $\pi_k = n/N$.

Result 1 With this procedure $\mathrm{E}(S_k^*|m) = 1$,

$$\mathrm{var}(S_k^*|m = r) = 1 - \pi_{k|r}, r = 0, 2, 3, \ldots, n,$$
$$\mathrm{var}(S_k^*|m = 1) = \frac{1 - \psi_k}{2} = 1 - \pi_{k|n-1} + \pi_{k|n-1} - \frac{1 + \psi_k}{2},$$

where $\pi_{k|r} = \mathrm{E}(X_k|m = r)$.

The proof is given in the Appendix.

We thus have $\mathrm{E}(S_k^*) = \mathrm{E}[\mathrm{E}(S_k^*|m)] = 1$ and

$$\begin{aligned}\mathrm{var}(S_k^*) &= \mathrm{Evar}(S_k^*|m) + \mathrm{varE}(S_k^*|m)\\ &= \mathrm{E}(1 - \pi_{k|r}) + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right)\Pr^*(m = n - 1)\\ &= 1 - \pi_k + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right)\Pr^*(m = n - 1).\end{aligned}$$

The variance of the diagonal is very slightly biased. Indeed

$$\mathrm{var}^*(S_k^*) = \check{\Delta}_{kk} + \left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right)\Pr^*(m = n - 1), k \in S.$$

The bias is small and often nonexistent. Indeed, $(m = n-1)$ is a rare event. Moreover, $\pi_{k|n-1} - (1 + \psi_k)/2$ is null except if one of the $\pi_{k|n-1}$ is smaller than $1/2$, which is also rare except in case where the sample size is very small. Moreover, we always have

$$\sum_{k\in S}\left(\pi_{k|n-1} - \frac{1 + \psi_k}{2}\right) = 0.$$

Below, the simulations will show that even for very small sample sizes, the bias is negligible.

The same results can be derived by taking ϕ_k in place of π_k. The D_{kk} can sometimes be larger than 1. In this case, we advocate to take $\phi_k = 0$. We can thus define two bootstrap methods depending on whether the inclusion probabilities are π_k or ϕ_k. The first case will be referred to as π-bootstrap and the second as ϕ-bootstrap. The bootstrap sample size always remains fixed, i.e.

$$\sum_{k\in S} S_k^* = n,$$

which implies that

$$\sum_{k\in S} \mathrm{E}^*(S_k^*) = n \quad \text{and} \quad \sum_{k\in S} \mathrm{cov}^*(S_k^*, S_\ell^*) = 0.$$

Unfortunately, we cannot show theoretically that (7) or (10) are satisfied, which are conditions for the validity of bootstrap variance estimation. However, the effect of ignoring (7) or (10) is not important in our empirical study. This is certainly due to the fact that the marginal constraints given in (13) and (14) are satisfied.

8 Simulation studies

8.1 Comparison with existing variance estimators for the total

In the first part of the simulation study, we examined the performance of the estimator using the proposed method and then compared this estimator with other variance estimators. We ran simulations on the MU284 population from Särndal et al. (1992) from where we selected samples of size $n = 2$, $n = 10$ and $n = 40$ with inclusion probabilities proportional to variable P75 (population in 1975). We used a maximum entropy design (also called conditional Poisson sampling) because this method maximizes the entropy of the sampling design subject to given inclusion probabilities and fixed sample size and can be implemented very quickly (see Tillé 2006). The variable of interest was RMT85 (revenues from 1985 municipal taxation). We compared the HT-estimator, the SYG-estimator, the H-estimator, the π-bootstrap and the ϕ-bootstrap of the variance for the total of RMT85. We ran 10,000 simulations and, in each of them, we used 10,000 bootstrap replications. Due to the simplicity of the method, the simulations were achieved in a few hours. Table 1 shows the relative bias given by

$$\mathrm{RB} = \frac{\mathrm{E}_{sim}[v_{boot}(\widehat{Y}^*)] - \mathrm{var}(\widehat{Y})}{\mathrm{var}(\widehat{Y})}$$

and the coefficients of variation given by

$$\mathrm{CV} = \frac{\sqrt{\mathrm{var}_{sim}[v_{boot}(\widehat{Y}^*)]}}{\mathrm{var}(\widehat{Y})}$$

of the HT-estimator, SYG-estimator and the Bootstrap estimator, where $\mathrm{E}_{sim}(.)$ and $\mathrm{var}_{sim}(.)$ respectively denote the expectation and the variance under the sample selection mechanism estimated by the simulations. Although only the HT-estimator and the SYG-estimator are strictly unbiased, the relative bias of the π-bootstrap method given by the simulations is still smaller. All the bias computed by simulation are nevertheless very small and are not significantly different from zero. The simulations also show that the HT-estimator is very unstable and that the bootstrap method performs as well as the SYG-estimator and the H-estimator. These simulations show that the bootstrap leads to an estimator of the variance that is at least as efficient as the SYG-estimator even for a very small sample size ($n = 2$).

Table 1 Relative bias and coefficients of variation of the HT-estimator, the SYG-estimator, H-estimator, the π-bootstrap and the ϕ-bootstrap

Estimator	Relative bias (%)	Coefficients of variation
$n = 2$		
HT-estimator	1.78511	1.98240
SYG-estimator	1.40424	1.80025
H-estimator	3.56754	1.84788
π-Bootstrap	3.77461	1.85225
ϕ-Bootstrap	1.14898	1.80014
$n = 10$		
HT-estimator	2.62246	1.31354
SYG-estimator	0.69995	0.50915
H-estimator	2.74196	0.53513
π-Bootstrap	0.76149	0.52311
ϕ-Bootstrap	0.30085	0.50914
$n = 40$		
HT-estimator	−1.53914	1.38550
SYG-estimator	−0.11598	0.26809
H-estimator	−0.35775	0.26119
π-Bootstrap	−0.19534	0.26211
ϕ-Bootstrap	−0.15363	0.26830

8.2 Performance in the case of variance estimation of other functions of interest

In the second part of the simulation study, we ran simulations in order to examine performance in relation to the variance of nonlinear functions of interest. Besides the total, the ratio of two totals, the median and the Gini index were also used as a function of interest. In the case of nonlinear statistics, the variances under the simulations, say the Monte Carlo variances were considered as the true variances of the estimators. A population of 150 units was generated from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, with $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0, 7)$, $\varepsilon_k \sim \mathcal{N}(0, 1)$ and $\sigma = 15$. The regression parameters were $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. The model and its parameters were chosen intentionally to have a distribution for y similar to a lognormal, as it is often used for income distributions, with a correlated and positive explanatory variable x in the regression model. From this population, 1000 samples were drawn using, as in the previous section, a maximum entropy sampling design with unequal inclusion probabilities. Concerning the inclusion probabilities, they were calculated proportional to the values of a variable z, which was generated from equation $z = y^{0.2} p$ where $p \sim \ln\mathcal{N}(0, 0.25)$. In this manner, the correlation between y and z is about 0.5. We knowingly used a large sample rate $n/N = 1/3$ and a skewed population in order to better illustrate the performance of the tested bootstrap methods. From each of these samples, we calculated four statistics: the total, the median, the Gini index of variable y and the ratio of total of variable y on the total of variable x.

From each of the 1,000 initial samples, 1,000 bootstrap samples were selected using three different bootstrap methods. Besides the new bootstrap method, two other resampling methods were tested. The first one is the generalization of the bootstrap method without replacement proposed by Booth et al. (1994) for unequal inclusion probabilities (Chauvet 2007). This bootstrap method of Booth et al. (1994) is itself a variant of the initial bootstrap with replacement method that consists of creating an artificial population from the initial sample and then drawing bootstrap samples from it with the same design as the initial one (Gross 1980; Chao and Lo 1985). After drawing the bootstrap samples, the estimators and their variances were computed for each of the initial samples and then the means of these variances were then compared with the approximations of the true variances. Note that the median is not a smooth function of the total. Estimating its variance can therefore be difficult, but the simulations show that in this case bootstrap methods perform well. The second one is the method proposed in Antal and Tillé (2011a).

In order to measure the performance of the new method and compare it with the other ones, the following five indicators were used:

- Lower error rate (L) in %

$$L = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} > \theta\right],$$

  where $I[a] = 1$ if a is true and $I[a] = 0$ elsewhere,
- Upper error rate (U) in %

$$U = \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} < \theta\right].$$

- Total error rate (ER) in %

$$ER = 100 - \frac{100}{sim}\sum_{i=1}^{sim} I\left[\widehat{\theta} - 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)} \leq \theta \leq \widehat{\theta} + 1.96 \times \sqrt{\text{var}(\widehat{\theta}^*)}\right].$$

- Relative Bias

$$RB = 100 \times \frac{\text{var}(\widehat{\theta}^*) - \text{var}_{sim}(\widehat{\theta})}{\text{var}_{sim}(\widehat{\theta})} = 100 \times \frac{B}{\text{var}_{sim}(\widehat{\theta})},$$

  where B is the Bias of the $\text{var}(\widehat{\theta}^*)$.
- Relative Root Mean Squared Error

$$RRMSE = 100 \times \frac{\sqrt{B^2 + \text{var}[\text{var}(\widehat{\theta}^*)\,]}}{\text{var}_{sim}(\widehat{\theta})}.$$

Table 2 Performance of the resampling methods in maximum entropy sampling design

UPWOR	L	U	ER	Relative bias	RRMSE
Total					
New method	1.1	6.4	7.5	0.1121	35.4938
Antal and Tillé method (2011a,b)	0.6	6.2	6.8	−1.3094	33.9250
Bootstrap WOR	1.1	6.9	8.0	−1.6084	34.7805
Median					
New method	4.1	6.9	11.0	−1.0564	58.8889
Antal and Tillé method (2011a,b)	4.1	7.3	11.4	−6.5615	50.6651
Bootstrap WOR	3.4	7.5	10.9	2.8753	61.9642
Gini					
New method	1.5	5.1	6.6	3.5753	39.4669
Antal and Tillé method (2011a,b)	2.8	6.2	9.0	−1.0795	35.4389
Bootstrap WOR	1.6	5.1	6.7	−1.0276	30.9325
Ratio					
New method	2.0	3.7	5.7	2.0403	41.1975
Antal and Tillé method (2011a,b)	2.0	3.6	5.6	−3.1038	40.1890
Bootstrap WOR	2.1	4.7	6.8	−2.8664	38.2802

The *RB* gives a measure of the bias of the estimator of variance. The *RRMSE* measures its accuracy and in the case of unbiasedness of the variance estimator it is equal to the variation coefficients. The *Error Rates* allow us to evaluate the capacity of the methods to provide a valid inference. The lower and the upper error rates give us an idea of how skewed the distribution of the estimator $\widehat{\theta}$ is.

Table 2 presents the results of the application of the resampling methods for a maximum entropy design with inclusion probabilities proportional to variable z. In the proposed bootstrap method, the Hájek approximation given in (4) is used, which gives us $\phi_k = \pi_k$. Each method provide confidence intervals around 93–94 % for the total, the ratio and the Gini index, and 89–90 % for the median. The column of the relative biases directly shows that, in each case of the four functions of interest, the methods perform well and give relative biases of 1–3 %.

Note that a high underestimation of the variance of a function of interest could result in a low coverage rate, and therefore high error rates for the function of interest, which is probably the case here. Regarding the relative root mean square errors, the same trend can be observed. The three treated method perform identically, giving a value of RRMSE around 30–40 % for the total, the Gini index and the ratio and 60 % for the median. In general, there is no major difference in performance between the proposed methods, the estimators are unbiased, or have a slight bias for each function. The RRMSE have the same order and the error rates show a slightly positively skewed distribution, with coverage rates between 90 and 95 %.

The new method thus provides essentially the same results as the other mentioned methods, but its application is simpler: it does not require a correction factor or artificial

population. Besides having at least the same performance as the method of artificial populations, its main advantage is that it is easy to implement and fast. Thus, the samples can be directly used to compute the variance of the functions of interest.

9 Discussion and interest of the method

The new method provides similar results as the Antal and Tillé (2011a) methods by using a mixture of several sampling designs. They both satisfy conditions (5) and (6) or (8) and (9). The proposed method is simpler because it is easier to implement. Mainly the doubled half bootstrap consists of selecting twice half the sample, which is particularly simple. Moreover the double half bootstrap limits the number of replication of the units in the bootstrap sample (maximun 3 and mainly 2).

If the S_k^* are not integer, they define a bootstrap weighting system. The interest of a bootstrap method that uses a discrete random variable S_k^* is that a bootstrap sample can be defined. Each unit is simply replicated S_k^* times. The units of the bootstrap samples have the same Horvitz–Thompson weights as in the original sample.

These bootstrap methods compare favorably with the best of the classical variance estimates for linear statistics, and also apply to nonlinear statistics. Its simplicity, its speed and its efficiency speak in its favour. The bootstrap sample does not need to be reweighted. There is no need for artificial populations and extreme samples are also avoided because the units can be repeated twice or rarely three times. The bootstrap samples can directly be used to provide estimates.

Acknowledgments We are grateful to a referee for his/her very pertinent comments that helped us to improve the quality of this manuscript. This research was performed when Erika Antal was research assistant at the University of Neuchâtel.

Appendix

Lemma 1 *If a sample S is selected by a Poisson sampling design with inclusion probabilities π_k in a population U of size N, if n_s denotes the random sample site, then*

$$\pi_{k|N-1} = \Pr(k \in S | n_S = N-1) = 1 - \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in U} \frac{1-\pi_\ell}{\pi_\ell}}.$$

Proof We have

$$\Pr(k \notin S \text{ and } n_S = N-1) = (1-\pi_k) \prod_{\ell \neq k} \pi_\ell = \frac{1-\pi_k}{\pi_k} \prod_{\ell \in U} \pi_\ell.$$

Thus

$$\Pr(n_S = N-1) = \sum_{k \in U} \Pr(k \notin \text{ and } n_S = N-1) = \sum_{k \in U} \frac{1-\pi_k}{\pi_k} \prod_{\ell \in U} \pi_\ell,$$

which gives the complementary of the conditional probability of Lemma 1.

$$\Pr(k \notin S | n_S = N-1) = \frac{\Pr(k \notin S \text{ and } n_S = N-1)}{\Pr(n_S = N-1)} = \frac{\frac{1-\pi_k}{\pi_k}}{\sum_{\ell \in U} \frac{1-\pi_\ell}{\pi_\ell}}.$$

Lemma 1. can also be derived from Expression (5.12) of Result 22 in Tillé (2006).

Proof of Result 1

Proof Let $\pi_{k|r} = \mathrm{E}(X_k | m = r)$. These conditional probabilities are not easy to compute. A recursive relation for computation is given for instance in Tillé (2006, p. 81). Fortunately, we do not have to compute this conditional expectation in order to proof the result except for case $r = n-1$. However we will use it in the following reasoning. We have

$$\Pr^*(S_k^* = j | m = r \quad \text{and} \quad n-r \text{ is even}) = \begin{cases} (1-\pi_{k|r})/2 & \text{if } j=0 \\ \pi_{k|r} & \text{if } j=1 \\ (1-\pi_{k|r})/2 & \text{if } j=2, \end{cases}$$

$$\Pr^*(S_k^* = j | m = r, n-r \text{ is odd, and } r < n-1) = \begin{cases} \left(1-\pi_{k|r}\right)\frac{2n-1}{4n} & \text{if } j=0 \\ \pi_{k|r} + \left(1-\pi_{k|r}\right)\frac{3}{4n} & \text{if } j=1 \\ \left(1-\pi_{k|r}\right)\frac{2n-3}{4n} & \text{if } j=2 \\ \left(1-\pi_{k|r}\right)\frac{1}{4n} & \text{if } j=3, \end{cases}$$

and

$$\Pr^*(S_k^* = j | m = n-1) = \begin{cases} (1-\psi_k)/4 & \text{if } j=0 \\ (1+\psi_k)/2 & \text{if } j=1 \\ (1-\psi_k)/4 & \text{if } j=2. \end{cases}$$

References

Antal E, Tillé Y (2011a) A direct bootstrap method for complex sampling designs from a finite population. J Am Stat Assoc 106:534–543

Antal E, Tillé Y (2011b) Simple random sampling with over-replacement. J Stat Plan Inference 141:597–601

Beaumont J-F, Patak Z (2012) On the generalized bootstrap for sample surveys with special attention to poisson sampling. Int Stat Rev 80(1):127–148

Bertail P, Combris P (1997) Bootstrap généralisé d'un sondage. Annales d'Economie et de Statistique 46:49–83

Booth JG, Butler RW, Hall P (1994) Bootstrap methods for finite populations. J Am Stat Assoc 89:1282–1289

Brewer KRW, Donadio ME (2003) The high entropy variance of the Horvitz–Thompson estimator. Surv Methodol 29:189–196

Brewer KRW, Hanif M (1983) Sampling with unequal probabilities. Springer, New York

Chao MT, Lo SH (1985) A bootstrap method for finite population. Sankhy$\bar{A}$ 47:399–405

Chauvet G (2007) Méthodes de bootstrap en population finie. PhD thesis, Université Rennes 2

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Gross ST (1980) Median estimation in sample surveys. In: ASA proceedings of the section on survey research methods. American Statistical Association, pp 181–184

Hájek J (1981) Sampling from a finite population. Marcel Dekker, New York

Henderson T (2006) Estimating the variance of the Horvitz–Thompson estimator. Master's thesis, School of Finance and Applied Statistics, The Australian National University
Holmberg A (1998) A bootstrap approach to probability proportional-to-size sampling. In: ASA proceedings of the section on survey research methods. American Statistical Association, pp 378–383
Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
Kuk AYC (1989) Double bootstrap estimation of variance under systematic sampling with probability proportional to size. J Stat Comput Simul 31:73–82
Lahiri P (2003) On the impact of bootstrap in survey sampling and small-area estimation. Stat Sci 18:199–210
Mac Carthy PJ, Snowden CB (1985) The bootstrap and finite population sampling. Public Health Service Publication, Technical report
Mason D, Newton MA (1992) A rank statistic approach to the consistency of a general bootstrap. Ann Stat 20:1611–1624
Matei A, Tillé Y (2005) Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. J Off Stat 21(4):543–570
Patak Z, Beaumont J-F (2009) Generalized bootstrap for prices surveys. In paper presented at the 57th Session of the International Statistical Institute, Durban, South-Africa
Preston J, Henderson T (2007) Replicate variance estimation and high entropy variance approximations. In Papers presented at the ICES-III, June 18–21, 2007, Montreal, QC, Canada
Rao JNK, Wu CFJ (1988) Resampling inference for complex survey data. J Am Stat Assoc 83:231–241
Rao JNK, Wu CFJ, Yue K (1992) Some recent work on resampling methods for complex surveys. Surv Methodol 18:209–217
Saigo H, Shao J, Sitter RR (2001) A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. Surv Methodol 27(2):189–196
Särndal C-E, Swensson B, Wretman JH (1992) Model assisted survey sampling. Springer, New York
Sen AR (1953) On the estimate of the variance in sampling with varying probabilities. J Indian Soc Agric Stat 5:119–127
Shao J, Tu D (1995) The jacknife and bootstrap. Sprinter, New York
Sitter RR (1992a) Comparing three bootstrap methods for survey data. Can J Stat 20:135–154
Sitter RR (1992b) A resampling procedure for complex survey data. J Am Stat Assoc 87:755–765
Tillé Y (2006) Sampling algorithms. Springer, New York
Yates F, Grundy PM (1953) Selection without replacement from within strata with probability proportional to size. J R Stat Soc B15:235–261