Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population

Erika Antal & Yves Tillé

To cite this article: Erika Antal & Yves Tillé (2011) A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population, Journal of the American Statistical Association, 106:494, 534-543, DOI: 10.1198/jasa.2011.tm09767

To link to this article: https://doi.org/10.1198/jasa.2011.tm09767

Published online: 24 Jan 2012.

A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population

Erika ANTAL and Yves TILLÉ

In complex designs, classical bootstrap methods result in a biased variance estimator when the sampling design is not taken into account. Resampled units are usually rescaled or weighted in order to achieve unbiasedness in the linear case. In the present article, we propose novel resampling methods that may be directly applied to variance estimation. These methods consist of selecting subsamples under a completely different sampling scheme from that which generated the original sample, which is composed of several sampling designs. In particular, a portion of the subsampled units is selected without replacement, while another is selected with replacement, thereby adjusting for the finite population setting. We show that these bootstrap estimators directly and precisely reproduce unbiased estimators of the variance in the linear case in a time-efficient manner, and eliminate the need for classical adjustment methods such as rescaling, correction factors, or artificial populations. Moreover, we show via simulation studies that our method is at least as efficient as those currently existing, which call for additional adjustment. This methodology can be applied to classical sampling designs, including simple random sampling with and without replacement, Poisson sampling, and unequal probability sampling with and without replacement.

KEY WORDS: One–one resampling design; Poisson sampling; Replications; Simple random sampling; Unequal probability sampling; Variance estimation.

1. INTRODUCTION

Resampling methods such as the bootstrap and jackknife are largely used to estimate variances across a broad spectrum of statistical contexts. In survey sampling, the variances of even simple estimators depend on the sampling design, and can take very complex forms, particularly when the sampling design is elaborate. The classical bootstrap method, developed by Efron (1979) cannot be directly applied to cases of sampling from a finite population because the identical and independent distribution assumption fails under sampling without replacement. Gross (1980) and Chao and Lo (1985) have proposed a method for variance estimation based on reconstructing artificial populations from the sample. Bootstrap samples are then selected from this artificial population using the original sampling scheme. Another important class of methods arises from the rescaled bootstrap (Rao and Wu 1988) which consists of modifying the sample values of the variable of interest to construct an unbiased estimator of the variance in the linear case. Other methods have also been proposed by McCarthy and Snowden (1985), Kuk (1989), Rao, Wu, and Yue (1992), Shao and Tu (1995), Sitter (1992a, 1992b), Booth, Butler, and Hall (1994), Holmberg (1998).

In this article, we propose a new methodology that can be applied to classical sampling designs both with and without replacement, as well as both equal and unequal inclusion probabilities. Our methodology consists of selecting bootstrap samples from the original sample in such a way that it eliminates the need for scaling, weighting of the sample, and using artificial populations. We argue that if the aim is variance estimation, the resampling design must be radically different from that which generates the original data. We then proceed to construct an ad hoc resampling design by mixing several designs, such that the bootstrap variance is equal to the estimator of the variance in the linear case, and such that the bootstrap sample has the same expected sample size as that of the actual sample size of the data, and can thus be treated as the original sample. This feature is particularly attractive because imputation, weighting for nonresponse, and calibration can thus be carried out without the need for any additional considerations or corrective techniques. In sampling without replacement, the main idea consists in selecting bootstrap samples by mixing sampling with and without replacement in order to reproduce a variance estimator that comprises the finite population correction.

The remainder of the article is structured as follows. We will first review basic notions of the theory of survey sampling, and provide an overview of the most frequently used sampling designs. We will then introduce two new sampling designs, simple random sampling with over-replacement and one–one resampling, that are used exclusively in resampling. We will then define sufficient conditions for a direct unbiased estimator for the variance of the total in resampling designs, and provide the construction of the algorithms used to draw such samples for several basic sampling designs. Finally, we supplement the theoretical proofs with results of simulation studies performed on several functions of interest, including the total, the median, the Gini index, and the ratio of totals. These results are compared to those obtained under resampling methods currently used, such as the classical bootstrap with and without replacement. We conclude with comparative remarks on our proposed methodology, and propose additional development and future research on the topic.

2. SAMPLING DESIGN AND ESTIMATION

Consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ and the variable of interest y that takes the value y_k on unit k, for all k in U. A first aim is to estimate the total of the interest variable: $Y = \sum_{k \in U} y_k$. A random sample is a random vector $\mathbf{S} = (S_1, \ldots, S_k, \ldots, S_N)'$, where S_k is the number of times

Erika Antal (E-mail: *erika.antal@unine.ch*) is Doctoral Student and Yves Tillé (E-mail: *yves.tille@unine.ch*) is Professor, Institute of Statistics, Faculty of Economics, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland. The authors would like to thank the Editor, associate editor, referees, and Anthea Monod for their insightful comments and propositions which have resulted in significant improvement of this article. This work was supported by Equal Opportunity Office of the University of Neuchâtel.

Journal of the American Statistical Association
June 2011, Vol. 106, No. 494, Theory and Methods
DOI: 10.1198/jasa.2011.tm09767

unit k is selected in the sample. If the sample is selected without replacement, then S_k can only take the values 0 and 1. If the sample has a fixed sample size n, then $\sum_{k\in U} S_k = n$.

Let π_k be the expectation of S_k, that is, $\pi_k = \mathrm{E}(S_k)$. The joint expectation of two units k and ℓ is $\pi_{k\ell} = \mathrm{E}(S_k S_\ell)$. Moreover, $\Delta_{k\ell} = \mathrm{cov}(S_k, S_\ell) = \pi_{k\ell} - \pi_k\pi_\ell$. If the sample is selected without replacement, π_k is the inclusion probability of unit k and $\pi_{k\ell}$ is the joint inclusion probability of unit k and ℓ.

If $\pi_k > 0$, for all $k \in U$, then the total Y can be estimated in an unbiased manner by using the Horvitz–Thompson estimator $\widehat{Y} = \sum_{k\in U} S_k y_k/\pi_k$. The variance of $\widehat{Y}$ is

$$\mathrm{var}(\widehat{Y}) = \sum_{k\in U}\sum_{\ell\in U} \frac{y_k y_\ell}{\pi_k\pi_\ell}\Delta_{k\ell}. \tag{1}$$

Theoretically, if $\pi_{k\ell} > 0$, for all $k \neq \ell \in U$, this variance can be estimated in an unbiased manner by

$$\widehat{\mathrm{var}}(\widehat{Y}) = \sum_{k\in U}\sum_{\ell\in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k\pi_\ell}\frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{2}$$

Nevertheless, this variance estimator is often very unstable. It can even take negative values. When the sampling design has a fixed sample size, then the variance can be written

$$\mathrm{var}(\widehat{Y}) = \frac{-1}{2}\sum_{k\in U}\sum_{\ell\in U}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \Delta_{k\ell},$$

and, if $\pi_{k\ell} > 0$, for all $k \neq \ell \in U$, can be estimated by the Yates–Grundy estimator of variance:

$$\widehat{\mathrm{var}}(\widehat{Y}) = \frac{-1}{2}\sum_{k\in U}\sum_{\ell\in U} S_k S_\ell\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2 \frac{\Delta_{k\ell}}{\pi_{k\ell}}. \tag{3}$$

Expression (3) holds for sampling with or without replacement and can also be written in the quadratic form

$$\widehat{\mathrm{var}}_D(\widehat{Y}) = \sum_{k\in U}\sum_{\ell\in U} \frac{S_k S_\ell y_k y_\ell}{\pi_k\pi_\ell} D_{k\ell}, \tag{4}$$

with

$$D_{k\ell} = \begin{cases} -\displaystyle\sum_{\substack{j\in U\\ j\neq k}} S_j\frac{\Delta_{kj}}{\pi_{kj}} & \text{if } k = \ell \\ \dfrac{\Delta_{k\ell}}{\pi_{k\ell}} & \text{if } k \neq \ell. \end{cases} \tag{5}$$

When the sampling design with or without replacement has a fixed sample size, estimator (3) must be preferred to estimator (2). We shall show below in Result 2 that the presentation of estimator (3) in a quadratic form is needed to construct a resampling method that produces an unbiased estimator.

3. BASIC SAMPLING DESIGNS

In a *Poisson sampling design* with inclusion probabilities π_k, the S_k are N independent Bernoulli random variables with parameter π_k. Thus, $\Delta_{k\ell} = \pi_k(1-\pi_k)$ if $k = \ell$ and 0, if not. So $\Delta_{k\ell}/\pi_{k\ell} = 1 - \pi_k$ if $k = \ell$ and 0 if not.

Simple random sampling with replacement is very common. The sampling design is given by $\Pr(\mathbf{S} = \mathbf{s}) = N^{-n}\binom{n}{s_1\cdots s_k\cdots s_N}$, for all $\mathbf{s} \in \mathcal{R}_n$, where $\mathcal{R}_n = \{\mathbf{s} \in \mathbb{N}^n | \sum_{k=1}^N s_k = n\}$. It follows that $\Delta_{k\ell} = -n(N-1)/\{N^2(N-1)\}$ when $k \neq \ell \in U$ and $\Delta_{kk} = n(N-1)/N^2$ when $k \in U$. Since the sample size is fixed, we can construct an unbiased estimator by using the quadratic form based on the $D_{k\ell}$, defined in Expression (5), $D_{k\ell} = -1/(n-1)$ when $k \neq \ell \in U$, and $D_{kk} = 1$ when $k \in U$, which gives

$$\widehat{\mathrm{var}}(\widehat{Y}) = \frac{N^2}{n}\frac{1}{n-1}\sum_{k\in U} S_k(y_k - \widehat{\overline{Y}})^2, \tag{6}$$

where $\widehat{\overline{Y}} = n^{-1}\sum_{k\in U} S_k y_k$.

Unequal probability with replacement with fixed sample size, is a generalization of simple random sampling with replacement to unequal probabilities of selection. The distribution of this sampling design is multinomial:

$$\Pr(\mathbf{S} = \mathbf{s}) = \binom{n}{s_1\cdots s_k\cdots s_N}^{-1}\prod_{k\in U}\left(\frac{\pi_k}{n}\right)^{s_k} \quad \text{for all } \mathbf{s} \in \mathcal{R}_n.$$

In unequal probability sampling with replacement,

$$\Delta_{k\ell} = \frac{n(N-1)}{N^2} \times \begin{cases} \pi_k\left(1 - \dfrac{\pi_k}{n}\right) & \text{if } k = \ell \\ -\dfrac{\pi_k\pi_\ell}{n} & \text{if } k \neq \ell. \end{cases}$$

In order to construct an unbiased estimator of the variance, we can use the $D_{k\ell}$ defined in Expression (5), and we get $D_{k\ell} = -1/(n-1)$ when $k \neq \ell \in U$, and $D_{kk} = 1$ when $k \in U$. Curiously, $D_{k\ell}$ does not depend on the π_k's of the sampling design and are the same as simple random sampling with replacement. The unbiased variance estimator (3) becomes

$$\widehat{\mathrm{var}}(\widehat{Y}) = \frac{n}{n-1}\sum_{k\in U} S_k\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}}{n}\right)^2. \tag{7}$$

Simple random sampling without replacement is defined by the following sampling design: $\Pr(\mathbf{S} = \mathbf{s}) = \binom{N}{n}^{-1}$, for all $\mathbf{s} \in \mathcal{S}_n$, where

$$\mathcal{S}_n = \left\{\mathbf{s} \in \{0,1\}^N \middle| \sum_{k=1}^N s_k = n.\right\}.$$

We thus have $\Delta_{k\ell} = -n(N-n)/\{N^2(N-1)\}$ when $k \neq \ell \in U$, $\Delta_{kk} = n(N-n)/N^2$ when $k \in U$, $\Delta_{k\ell}/\pi_{k\ell} = -(N-n)/\{N(n-1)\}$ when $k \neq \ell \in U$, and $\Delta_{kk}/\pi_{kk} = (N-n)/N$ when $k \in U$.

Unequal probability sampling without replacement and with fixed sample size is much more complex. The first problem is that there are many methods of sampling without replacement and with unequal probabilities. Each method provides a specific matrix of joint inclusion probabilities. These inclusion probabilities are, however, very similar if the sampling has a large entropy (Berger 1998; Brewer and Donadio 2003; Henderson 2006), such as the random systematic design (Madow 1949) or the Rao–Sampford design (Rao 1965; Sampford 1967), the Brewer design (Brewer 1975), the maximum entropy design or the random pivotal design (Tillé 2006, pp. 79–95 and p. 106). The second problem is that these inclusion probabilities can never be simplified. So, a simpler expression of variance than (1) and its estimator (2) cannot be constructed. Several approximations of variance based on a simple sum have been proposed, however. These approximations are obviously biased, but simulations have shown that they have

smaller mean squared errors than estimators (2) and (4) (Hájek 1981; Matei and Tillé 2005). There are thus various ways to estimate the variance. The strictly unbiased estimator consists of computing the $D_{k\ell}$ by expression (5). A general biased and simple estimator of variance is given by

$$\widehat{\mathrm{var}}(\widehat{Y}) = \sum_{k\in S} c_k \left(\frac{y_k}{\pi_k} - \frac{\sum_{k\in S} c_k y_k/\pi_k}{\sum_{k\in S} c_k} \right)^2,$$

where the c_k are weights that we discuss further. This expression can be viewed as an approximation of the $D_{k\ell}$ given in expression (5), by

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

Diverse values have been proposed for the c_k:

1. A simple value was given by Hájek (1981), who proposed using
$$c_{k1} = \frac{n}{n-1}(1-\pi_k). \tag{8}$$
2. Deville and Tillé (2005) proposed c_k such that
$$c_{k2} - \frac{c_{k2}^2}{\sum_{j\in U} S_j c_{j2}} = 1 - \pi_k. \tag{9}$$
In this case, the diagonal elements $\widetilde{D}_{kk2}$ of the approximated matrix are equal to $1-\pi_k$. A solution does not always exist for this equation, for instance when $n = 2$.
3. One could also take the c_k such that
$$c_{k3} - \frac{c_{k3}^2}{\sum_{j\in U} S_j c_{kj}} = -\sum_{\substack{j\in U \\ j\neq k}} S_j \frac{\Delta_{kj}}{\pi_{kj}}, \tag{10}$$
but this approximation needs to solve a nonlinear system of equations. In this case, the diagonal elements of the approximated matrix
$$\widetilde{D}_{kk3} = -\sum_{\substack{j\in U \\ j\neq k}} S_j \frac{\Delta_{kj}}{\pi_{kj}}$$
and correspond to the diagonal of the matrix used for the Yates–Grundy estimator of variance given in (5).

Simple random sampling with over-replacement was recently proposed by Antal and Tillé (2010). The sampling design is defined by $\Pr(S_1 = x_1, \ldots, S_N = x_N) = (\mathrm{card}\,\mathcal{R}_n)^{-1} = \binom{N+n-1}{n}^{-1}$. The $\binom{N+n-1}{n}$ samples with replacement have exactly the same probability of being selected. The marginal distribution of S_k is given by

$$\Pr(S_k = j) = \binom{N+n-1}{n}^{-1} \times \binom{N-1+n-j-1}{n-j}, \qquad j = 0, \ldots, n,$$

which is an inverse hypergeometric distribution.The expectation is $\mathrm{E}(S_k) = n/N$, and the matrix of $\Delta_{k\ell}$ is given by

$$\Delta_{k\ell} = \frac{(N-1)(N+n)n}{N^2(N+1)} \times \begin{cases} 1 & \text{if } k = \ell \\ -\dfrac{1}{N-1} & \text{if } k \neq \ell. \end{cases}$$

This design has a larger variance than sampling with replacement and will be used only for resampling.

4. RESAMPLING AND SUFFICIENT CONDITIONS

Define the random set S that contains the list of labels for the units selected in the sample $\mathbf{S}$. If a unit is selected several times in the sample, the labels can appear several times in S. For instance, if from population $U = \{1, 2, 3, 4, 5, 6\}$, we select a sample $\mathbf{S}$ that takes the value $(0, 2, 1, 0, 3, 1)$, the set S takes the value $\{2, 2, 3, 5, 5, 5, 6\}$. A resampling method is a second stage on sampling from sample S. A subsample $\mathbf{S}^* = (S_k^*, k \in S)$ can thus be presented as a sequence of discrete nonnegative random variables S_k^* that denote the number of times unit k is resampled. For example, if, in the above example, $\mathbf{S}^*$ takes the values $(1, 0, 3, 0, 2, 0, 1)$, then the subsample set S^* will be $\{2, 3, 3, 3, 5, 5, 6\}$. The S_k^* are generally not independent. A correlation is indeed necessary to obtain an unbiased estimation of the variance when the sample size is fixed. The resampling sample size is denoted by n^*.

In fact, a resampling method is a second phase of sampling that can depend on the first phase. Let $\mathrm{E}^*(\cdot) = \mathrm{E}(\cdot|S)$, $\mathrm{var}^*(\cdot) = \mathrm{var}(\cdot|S)$ and $\mathrm{cov}^*(\cdot,\cdot) = \mathrm{cov}(\cdot,\cdot|S)$ denote, respectively, the conditional expectation, variance and covariance under the resampling design with respect to the original design. Moreover, let $\Pr^*(\cdot) = \Pr(\cdot|S)$ denote the probability under the resampling design and conditionally to the original design. Let $\alpha_k = \mathrm{E}^*(S_k^*)$, $\alpha_{k\ell} = \mathrm{E}^*(S_k^* S_\ell^*)$ and $\mathrm{cov}^*(S_k^*, S_\ell^*) = \Sigma_{k\ell} = \alpha_{k\ell} - \alpha_k\alpha_\ell$. The resampled estimator of the total is defined as $\widehat{Y}^* = \sum_{k\in S} y_k S_k^*/\pi_k$. This estimator is generally biased, its conditional expectation is

$$\mathrm{E}^*(\widehat{Y}^*) = \sum_{k\in S} \frac{y_k \mathrm{E}^*(S_k^*)}{\pi_k} = \sum_{k\in S} \frac{y_k \alpha_k}{\pi_k}. \tag{11}$$

Note that α_k can depend on S. If $\alpha_k = 1$, then the estimator is unbiased.

The conditional variance of the resampled estimator is

$$\mathrm{var}^*(\widehat{Y}^*) = \sum_{k\in S}\sum_{\ell\in S} \frac{y_k y_\ell}{\pi_k \pi_\ell} \Sigma_{k\ell}. \tag{12}$$

This directly leads to two fundamental results:

Result 1. A sufficient condition for $\mathrm{E}^*(\widehat{Y}^*) = \widehat{Y}$ is $\alpha_k = 1$, for all $k \in U$.

This result directly comes from the equality between Expression (11) and the Horvitz–Thompson estimator.

Result 2. A sufficient condition for $\mathrm{var}^*(\widehat{Y}) = \widehat{\mathrm{var}}(\widehat{Y})$, is $\Sigma_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$, for all $k, \ell \in U$ and a sufficient condition for $\mathrm{var}^*(\widehat{Y}) = \widehat{\mathrm{var}}_D(\widehat{Y})$, is $\Sigma_{k\ell} = D_{k\ell}$, for all $k, \ell \in U$ if the sample size is fixed.

This result directly comes from the equality between Expressions (2) and (12) or between Expressions (4) and (12) when the sample size of S is fixed.

In fact, the main idea of this article is to develop resampling methods that satisfy conditions given in Results 1 and 2. This idea leads us to choose a sampling design for S^* that is completely different from the sampling design used for S. Indeed, Result 1 is generally not satisfied by using the same sampling design for S and S^*, because $\Sigma_{k\ell}$ and $D_{k\ell}$ are of very different natures: the $\Sigma_{k\ell}$ are variances and covariances, but the $D_{k\ell}$ are not.

Let $\widehat{\theta}$ be an estimator of a function of interest θ. Estimator $\widehat{\theta}$ is a function of the observed data $\{(y_k, \pi_k), k \in S\}$. The bootstrap estimator $\widehat{\theta}^*$ is the same function as $\widehat{\theta}$, applied on the bootstrap data $\{(y_k, \pi_k), k \in S^*\}$. Practically, a sequence of bootstrap samples $\mathbf{S}^{*1}, \ldots, \mathbf{S}^{*m}$ are selected with the bootstrap design. The bootstrap variance given in (12) is approximated by

$$\widetilde{\text{var}}^*(\widehat{\theta}^*) = \frac{1}{m-1}\sum_{j=1}^{m}(\widehat{\theta}_j^* - \overline{\widehat{\theta}})^2,$$

where $\widehat{\theta}_j^*$ is the bootstrap estimator computed on the jth bootstrap sample and $\overline{\widehat{\theta}} = (1/m)\sum_{j=1}^{m}\widehat{\theta}_j^*$.

5. THE SIMPLEST EXAMPLE: RESAMPLING FROM A POISSON SAMPLE

In a Poisson sampling design, $\Delta_{k\ell}/\pi_{k\ell} = 0$ when $k \neq \ell \in U$, and $\Delta_{kk}/\pi_{kk} = 1 - \pi_k$ when $k \in U$. The resampling design must be such that $\text{E}^*(S_k^*) = 1$, $\text{var}^*(S_k^*) = 1 - \pi_k$, and $\text{cov}^*(S_k^*, S_\ell^*) = 0$, for all $k \neq \ell$. Algorithm 1 can be used to generate such S_k^*'s. The main idea consists of selecting a part of the units without replacement and a part with replacement with Poisson random variables in order to reproduce the finite population correction $1 - \pi_k$.

With Algorithm 1, the expectations, variances and covariances of the S_k^* can be computed $\text{E}^*(S_k^*) = \text{E}^*(S_{kA}^*) + \text{E}^*(S_{kB}^*) = \pi_k + 1 \times (1 - \pi_k) = 1$. Moreover, $\text{var}^*(S_k^*) = \text{E}^*[\text{var}^*(S_k^*|S_{kA}^*)] + \text{var}^*[\text{E}^*(S_k^*|S_{kA}^*)] = 1 - \pi_k$. This bootstrap method provides the exact Horvitz–Thompson estimator in the linear case. Indeed, $\text{var}^*(\widehat{Y}^*) = \text{var}^*(\sum_{k\in S} y_k S_k^*/\pi_k) = \sum_{k\in S} y_k^2(1-\pi_k)/\pi_k^2 = \widehat{\text{var}}(\widehat{Y})$.

6. THE ONE–ONE RESAMPLING DESIGN

The one–one design is a sampling design defined only for resampling. It is an ad hoc construction used to randomly select n units from a sample of size n in such a way that the expectation and the variance of S_k^* are equal to 1, that is, $\text{E}^*(S_k^*) = 1$ and $\text{var}^*(S_k^*) = 1$. This sampling design is a mixture between a simple random sampling with replacement and a simple random sampling with over-replacement. Its implementation is given in Algorithm 2.

Algorithm 1 Resampling procedure for Poisson sampling

Define, independently, for $k \in S$:

- S_{kA}^* is a Bernoulli random variable with parameter π_k.
- If $S_{kA}^* = 1$ then $S_{kB}^* = 0$,
  else S_{kB}^* is a Poisson random variable with parameter $\lambda = 1$.
- The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$.

Algorithm 2 The one–one resampling design

- If $n = 2$, then

$$S_1^* = \begin{cases} 0 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/2 \end{cases}$$

  and $S_2^* = 2 - S_1^*$.
- If $n \geq 3$, then
  - Compute:

$$m = \left\lfloor \frac{1}{2}\left(1 + \sqrt{\frac{4n^2+5n-1}{n-1}}\right)\right\rfloor, \tag{13}$$

    where $\lfloor x \rfloor$ is the largest integer less than or equal to x and

$$\alpha = \frac{m(n-1)(m+1) - n(n+1)}{2m(n-1)}. \tag{14}$$

  - Define the random variable

$$\tilde{n} = \begin{cases} m & \text{with a probability } \alpha \\ m+1 & \text{with a probability } 1-\alpha. \end{cases}$$

  - Select a simple random sample with overreplacement with sample size $\tilde{n}$ from S. This sample is denoted by S_{kA}^*.
  - Select a simple random sample with replacement with sample size $n - \tilde{n}$ from S. This sample is denoted by S_{kB}^*. This second sample is independent from the first one.
  - The final sample is $S_k^* = S_{kA}^* + S_{kB}^*$.

Result 3. If S_k^* is the number of times unit k is selected by the one–one resampling design described in Algorithm 2, then $\text{E}^*(S_k^*) = 1, \text{var}^*(S_k^*) = 1, \text{cov}^*(S_k^*, S_\ell^*) = -1/(n-1)$, for all $k \neq \ell$.

Proof. The case where $n = 2$ is obvious. For the case where $n \geq 3$, we have that $\text{E}^*(S_k^*|\tilde{n}) = \text{E}^*(S_{kA}^*|\tilde{n}) + \text{E}^*(S_{kB}^*|\tilde{n}) = \tilde{n}/n + (n-\tilde{n})/n = 1$. Thus $\text{E}^*(S_k^*) = \text{E}^*\text{E}^*(S_k^*|\tilde{n}) = 1$. Moreover,

$$\begin{aligned}\text{var}^*(S_k^*|\tilde{n}) &= \text{var}^*(S_{kA}^*|\tilde{n}) + \text{var}^*(S_{kB}^*|\tilde{n}) \\ &= \frac{(n-1)(n+\tilde{n})\tilde{n}}{n^2(n+1)} + \frac{(n-\tilde{n})(n-1)}{n^2} \\ &= \frac{n-1}{n^2}\left[\frac{(n+\tilde{n})\tilde{n} + (n+1)(n-\tilde{n})}{(n+1)}\right].\end{aligned}$$

Since $\text{E}^*(S_k^*|\tilde{n}) = 1$,

$$\begin{aligned}\text{var}^*(S_k^*) &= \text{E}^*\,\text{var}^*(S_k^*|\tilde{n}) \\ &= \alpha\frac{n-1}{n^2}\left[\frac{(n+m)m + (n+1)(n-m)}{(n+1)}\right] \\ &\quad + (1-\alpha)\frac{n-1}{n^2} \\ &\quad \times \left[\frac{(n+m+1)(m+1) + (n+1)(n-m+1)}{(n+1)}\right] \\ &= \frac{(n-1)[n+n^2+m(1-2\alpha+m)]}{n^2(1+n)}.\end{aligned} \tag{15}$$

By plugging the value of α given in (14) and the value of m given in (13) in Expression (15), we get $\text{var}^*(S_k^*) = 1$. This sampling design has a fixed sample size, which implies that $\sum_{k\in S}\text{cov}^*(S_k^*, S_\ell^*) = \text{cov}^*(n, S_\ell^*) = 0$. Moreover, since all the units are treated symmetrically $\text{cov}^*(S_k^*, S_\ell^*) = -\text{var}^*(S_k^*)/(n-1)$. We thus have $\Sigma_{k\ell} = -1/(n-1)$ when $k \neq \ell \in U$ and $\Sigma_{kk} = 1$ for $k \in U$.

7. RESAMPLING FROM A SIMPLE RANDOM SAMPLE WITH REPLACEMENT

7.1 The Usual Bootstrap With Replacement

If the sample S is selected by means of simple random sampling with replacement, the formula of the estimated variance of the total estimator is already given in Expression (6). The usual bootstrap consists of selecting a sample from S with the same sampling design, that is, a simple random sampling design with replacement from S. In this case, the variance of the resampled estimator is $\text{var}^*(\widehat{Y}^*) = (N^2/n^2)\sum_{k\in S}(y_k - \overline{Y})^2$. The bootstrap variance slightly underestimates the unbiased estimator given in Expression (6). Indeed, $\widehat{\text{var}}(\widehat{Y}) = n/(n-1) \times \text{var}^*(\widehat{Y}^*)$. Actually, this underestimation is not very important if the sample size is large but can create problems if the samples are selected in strata with small sample sizes. Obviously, a correction factor can be applied in each stratum, but these procedures require a particular treatment of the bootstrap sample in each stratum.

7.2 Bootstrap by Using the One–One Sampling Design

The one–one simple random sampling design allows us to avoid the use of correction factors for the variance. Indeed, if the bootstrap sample is selected by a one–one design then the bootstrap variance is $\text{var}^*(\widehat{Y}^*) = [N^2/\{n(n-1)\}]\sum_{k\in S}(y_k - \overline{Y})^2$. In a one–one simple random sampling, the repetition of the units is slightly larger than with simple random sampling with replacement, which increases the variance by a factor of $n/(n-1)$. It is thus no longer necessary to multiply the bootstrap variance by this factor.

8. RESAMPLING FROM A SAMPLE SELECTED WITH UNEQUAL PROBABILITIES WITH REPLACEMENT

8.1 The Usual Bootstrap With Replacement

If the sample is selected with unequal probabilities, with replacement and with fixed sample size, the estimator of variance is given in (7). In this case, the matrix of $D_{k\ell}$ given in (5) does not depend on the π_k's, which means that the resampling design must be done with equal selection probabilities. A usual design consists of resampling by means of simple random sampling with replacement, which gives the bootstrap variance

$$\text{var}^*(\widehat{Y}^*) = \sum_{k\in S}\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}}{n}\right)^2.$$

With simple random sampling with replacement, the bootstrap variance thus suffers from a small underestimation. This problem can be annoying when the sample size is small and can be fixed by using a one–one simple random sampling.

8.2 Bootstrap by Using the One–One Sampling Design

If the bootstrap samples are selected with a one–one design, the bootstrap variance becomes

$$\text{var}^*(\widehat{Y}^*) = \frac{n}{n-1}\sum_{k\in S}\left(\frac{y_k}{\pi_k} - \frac{\widehat{Y}}{n}\right)^2,$$

and is exactly equal to the estimator of variance (7). The one–one design is thus a convenient design for resampling from a sample selected with unequal probabilities with replacement, particularly when the sample size is small.

9. RESAMPLING FROM A SIMPLE RANDOM SAMPLE SELECTED WITHOUT REPLACEMENT

9.1 Resampling Using Simple Random Sampling With Replacement

In simple random sampling without replacement, the estimator of variance is

$$\widehat{\text{var}}(\widehat{Y}) = \frac{N^2(N-n)}{nN}\frac{1}{n-1}\sum_{k\in S}(y_k - \widehat{\overline{Y}})^2. \tag{16}$$

A simple way of resampling consists of using a simple random sampling with replacement as a resampling design. In this case, $\text{var}^*(\widehat{Y}^*) = (N^2/n^2)\sum_{k\in S}(y_k - \overline{Y})^2$. Obviously, the bootstrap variance is not equal to the variance estimator. Indeed, $\widehat{\text{var}}(\widehat{Y}) = \text{var}^*(\widehat{Y}^*)(N-n)n/\{N(n-1)\}$, which means that the resampling variance does not take into account the loss of one degree of freedom and the finite population correction. The bootstrap variance must be corrected by a factor. This correction can become intricate if a large number of samples are selected in strata.

9.2 Resampling Using a With Replacement and a One–One Design

In order to avoid the use of a correction factor, one can use a mixture of a simple sampling without replacement and a one–one design as described in Algorithm 3 in order to directly reproduce the unbiased estimator of variance for the totals.

Result 4 gives the properties of Algorithm 3.

Result 4. If Algorithm 3 is used for the resampling design, (i) $\text{E}^*(S_k^*) = 1$, (ii) $\text{var}^*(S_k^*) = (N-n)/N$, (iii) $\text{cov}^*(S_k^*, S_\ell^*) = -(N-n)/\{N(n-1)\}$.

Proof. The case where $n - n^2/N < 2$ is trivial. For the case where $n - n^2/N \geq 2$, the expectation is given by $\text{E}^*(S_k^*) = \text{E}^*(S_{kA}^*) + \text{E}^*(S_{kB}^*|S_{kA}^*)\Pr^*(S_{kA} = 0) = n/N + 1 \times (1 - n/N) = 1$. Next, the variance is $\text{var}^*(S_k^*) = \text{E}^*[\text{var}^*(S_k^*|S_{kA}^*)] + \text{var}^*[\text{E}^*(S_k^*|S_{kA}^*)] = \text{var}^*(S_{kB}^*|S_{kA}^* = 0)\Pr^*(S_{kA}^* = 0) = 1 - n/N$. Finally, the covariances can be derived from the symmetry of treatment of the units, which implies that $\text{cov}^*(S_k^*|S_{kA}^*, S_{\ell A}^*) = -\text{var}^*(S_k^*)/(n-1) = -(N-n)/\{N(n-1)\}$.

If Algorithm 3 is used, the resampling variance is thus

$$\text{var}(\widehat{Y}^*) = N^2\frac{N-n}{nN}\frac{1}{n-1}\sum_{k\in S}(y_k - \widehat{\overline{Y}})^2,$$

and is exactly equal to the estimator of variance given in Expression (16).

Algorithm 3 Resampling using a with replacement and a one–one design

- If $n - n^2/N < 2$:
  - With a probability $q = n(N-n)/(2N)$, select randomly without replacement and with equal probabilities two units in S denoted by i and j. Next define $S_i^* = 2, S_j^* = 0, S_k = 1$, for all $k \notin \{i, j\}$.
  - With a probability $1 - q$, $S_k^* = 1$, for all $k \in S$.
- If $n - n^2/N \geq 2$:
  - Define

$$m = \begin{cases} \left\lfloor \dfrac{n^2}{N} \right\rfloor & \text{with probability } q \\ \left\lfloor \dfrac{n^2}{N} \right\rfloor + 1 & \text{with probability } 1-q, \end{cases}$$

    where $q = \lfloor n^2/N \rfloor + 1 - n^2/N$.
  - Select a sample S_{kA}^* from S with simple random sampling design without replacement with a sample size m.
  - From the set of units of S such that $S_{kA}^* = 0$, select a sample S_{kB}^* according to a one–one design, so S_{kB}^* has size $n - m$.
  - The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$.

10. RESAMPLING FROM A SAMPLE SELECTED WITH UNEQUAL PROBABILITIES WITHOUT REPLACEMENT

Unequal probability without replacement is obviously a more complicated problem. The main reason is that the unbiased estimators given in (2) and (4) of the variance can never been simplified, which makes it necessary to compute all the joint inclusion probabilities to estimate the variance. When the entropy of the sampling design is large, biased estimators given in (8), (9), and (10) have a smaller mean square error than estimators (2) and (4) (see Matei and Tillé 2005). For this reason, we do not propose using a bootstrap method that exactly reproduces the estimator of variance, but rather one that gives one of the three approximations. These methods are described in Algorithms 4 and 5.

Result 5 gives the properties of Algorithm 4.

Result 5. If Algorithm 4 is used for the resampling design, (i) $\mathrm{E}^*(S_k^*) = 1$, (ii) $\mathrm{var}^*(S_k^*) = 1 - \phi_k$, (iii) $\mathrm{cov}^*(S_k^*, S_\ell^*) = -q(1 - \phi_{k1} - \phi_{\ell 1} + \phi_{k\ell 1})/(n - m_1 - 1) - (1-q)(1 - \phi_{k2} - \phi_{\ell 2} + \phi_{k\ell 2})/(n - m_2 - 1)$.

Proof. First, the conditional expectation is given by $\mathrm{E}^*(S_k^*|m_j) = \mathrm{E}^*(S_{kA}^*|m_j) + \mathrm{E}^*(S_{kB}^*|m_j) = \phi_{kj} + 1 \times (1 - \phi_{kj}) = 1$. Thus, $\mathrm{E}^*(S_k^*) = \mathrm{E}^*\mathrm{E}^*(S_k^*|m) = 1$. Next, the conditional variance is $\mathrm{var}^*(S_k^*|m_j) = \mathrm{E}^*[\mathrm{var}^*(S_k^*|S_{kA}^*, m_j)|m_j] + \mathrm{var}^*[\mathrm{E}^*(S_k^*|S_{kA}^*, m_j), m_j] = \mathrm{var}^*(S_{kB}^*|S_{kA}^* = 0, m_j)\Pr^*(S_{kA}^* = 0|m_j) = 1 - \phi_{kj}, j = 1, 2$. Thus, $\mathrm{var}^*(S_k^*) = \mathrm{E}^*\,\mathrm{var}^*(S_k^*|m) + \mathrm{var}^*\,\mathrm{E}^*(S_k^*|m) = q(1 - \phi_{k1}) + (1-q)(1 - \phi_{k2}) = 1 - \phi_k$. Finally, the covariance is given by

$$\begin{aligned} &\mathrm{cov}^*(S_k^*, S_\ell^*|m_j) \\ &\quad = \mathrm{cov}^*\left[\mathrm{E}^*(S_k^*|S_{kA}^*, S_{\ell A}^*, m_j), \mathrm{E}^*(S_\ell^*|S_{kA}^*, S_{\ell A}^*, m_j)|m_j\right] \\ &\qquad + \mathrm{E}^*[\mathrm{cov}^*(S_k^*, S_\ell^*|S_{kA}^*, S_{\ell A}^*, m_j)|m_j] \\ &\quad = \mathrm{cov}^*(S_k^*, S_\ell^*|S_{kA}^* = 0, S_{\ell A}^* = 0, m_j) \\ &\qquad \times \Pr^*(S_{kA}^* = 0, S_{\ell A}^* = 0|m_j) \\ &\quad = -\frac{1}{n - m_j - 1} \times (1 - \phi_{kj} - \phi_{\ell j} + \phi_{k\ell j}). \end{aligned}$$

Thus

$$\begin{aligned} \mathrm{cov}^*(S_k^*, S_\ell^*) &= \mathrm{E}^*\,\mathrm{cov}^*(S_k^*, S_\ell^*|m) + \mathrm{cov}^*[\mathrm{E}^*(S_k^*|m), \mathrm{E}^*(S_\ell^*|m)] \\ &= -\frac{q}{n - m_1 - 1} \times (1 - \phi_{k1} - \phi_{\ell 1} + \phi_{k\ell 1}) \\ &\quad - \frac{1-q}{n - m_2 - 1} \times (1 - \phi_{k2} - \phi_{\ell 2} + \phi_{k\ell 2}). \end{aligned}$$

We have seen that according to the definition of the c_k, there are several ways to approximate the matrix of $D_{k\ell}$ by a matrix of $\widetilde{D}_{k\ell}$. The values of ϕ_k that reconstruct as best as possible the three approximations for c_k given in (8), (9), and (10) can be chosen by taking $1 - \phi_k = \widetilde{D}_{kk}$. Obviously, these resampling

Algorithm 4 Resampling for unequal probability sampling without replacement: Case 1

Case 1: $n - \sum_{k \in S} \phi_k \geq 2$.

- Select a sample S_{kA}^* without replacement with unequal inclusion probabilities ϕ_k (the choice of ϕ_k is discussed below) and fixed sample size. This sampling design is the same as the original design. If $n^* = \sum_{k \in S} \phi_k$ is not an integer, then define

$$m = \begin{cases} m_1 = \lfloor n^* \rfloor & \text{with probability } q \\ m_2 = \lfloor n^* \rfloor + 1 & \text{with probability } 1-q, \end{cases}$$

  where $q = \lfloor n^* \rfloor + 1 - n^*$. Also define ϕ_{k1} and ϕ_{k2} as the inclusion probabilities such that

$$\sum_{k \in S} \phi_{k1} = m_1, \qquad \sum_{k \in S} \phi_{k2} = m_2,$$

$$q\phi_{k1} + (1-q)\phi_{k2} = \phi_k \quad \text{for all } k \in S.$$

  Let $\phi_{k\ell 1}$ and $\phi_{k\ell 2}$ also be the joint inclusion probabilities of the design where sample sizes m_1 or m_2 were selected.
- From the set of units of S such that $S_{kA}^* = 0$, select a sample S_{kB}^* according to a one–one design.
- The resampling design is $S_k^* = S_{kA}^* + S_{kB}^*$.

Algorithm 5 Resampling for unequal probability sampling without replacement: Case 2

Case 2: $n - \sum_{k \in S} \phi_k < 2$.

- Compute $(\psi_k, k \in S) = 1 - H(1 - \phi_k, k \in S; 2)$ and $q = (n - \sum_{k \in S} \phi_k)/2$.
- With a probability q select a sample without replacement denoted by S_{kA}^* of size $n - 2$ from S by using inclusion probabilities ψ_k. Let $\psi_{k\ell}$ denote the joint inclusion probability of this design. From the two remaining units, select a one–one design denoted by S_{kB}^*. The final sample is $S_{kA}^* + S_{kB}^*$.
- With a probability $1 - q$, $S_k^* = 1$, for all k in S.

variances are not exactly equal to the estimator of variance, but they take into account the correction for finite population. Moreover, the diagonal terms are exactly the same as usual estimators of variance.

The case where $n - \sum_{k\in s} \phi_k < 2$ must also be treated. Consider the procedure used to compute the inclusion probabilities from a vector of positive values x_k. First, compute the quantities

$$\frac{nx_k}{\sum_{\ell\in U} x_\ell}, \tag{17}$$

$k = 1, \ldots, N$. For units for which these quantities are larger than 1, set $\pi_k = 1$. Next, the quantities are recalculated using (17) 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 $H(x_1, \ldots, x_N; n)$ denote the function that allows us to construct these inclusion probabilities from a vector of positive values $(x_1, \ldots, x_N)$. Function $H(\cdot; \cdot)$ allows us to define Algorithm 5 in order to select the bootstrap sample in the case where $n - \sum_{k\in s} \phi_k < 2$.

With Algorithm 5, $\mathrm{E}^*(S_k^*) = 1$ and

$$\mathrm{var}^*(S_k^*) = \frac{(1 - \psi_k)(n - \sum_{k\in S} \phi_k)}{2}$$

and

$$\mathrm{cov}^*(S_k^*, S_\ell^*) = \frac{(1 - \psi_k - \psi_\ell + \psi_{k\ell})(n - \sum_{k\in S} \phi_k)}{2}.$$

The ϕ_k can be chosen according to the three approximations given above in (8), (9), and (10).

11. MONTE CARLO SIMULATION STUDY FOR NUMERICAL COMPARISONS

First, we developed simulations for matrix reconstruction in order to confirm the theoretical results obtained in Section 10 on the new bootstrap methods for unequal probability sampling. As seen earlier, we distinguished the two cases depending on whether $n - \sum_{k\in S} \phi_k$ is greater than or equal to 2 or less than 2. We generated a population for each of these cases. We computed the matrices of Horvitz–Thompson and of Yates–Grundy variance estimators, as well as their approximations, we then ran sets of simulations to obtain the matrices of the variances using the new bootstrap method. We noticed that these matrices were very close to the respective approximations, so the method should provide estimators of variance that are very similar to the estimators given by the approximations. In order to be concise, we do not include the results of these simulations in this article.

Secondly, we also ran a set of simulations for the variance estimators in different sampling designs. In each case a population of 150 units was generated from the model $y_k = (\beta_0 + \beta_1 x_k^{1.2} + \sigma\varepsilon_k)^2 + c$, whit $x_k = |i_k|$ and $i_k \sim \mathcal{N}(0, 7)$, $\varepsilon_k \sim \mathcal{N}(0, 1)$ and $\sigma = 15$. The regression parameters are $\beta_0 = 12.5$, $\beta_1 = 3$ 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 with a sample size $n = 50$. 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.

Three sampling designs were tested: Poisson sampling, simple random sampling without replacement and a maximum entropy 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. In the case where the total was the function of interest, the goal was to reproduce the estimator of variance of the total. In fact, for the estimation of the total, estimators of variance can directly be computed. A resampling method is thus not necessary. However, simulations were also run in this case in order to test the performance of the methods.

From each of the 1000 initial samples, 1000 bootstrap samples were selected by means of five different bootstrap methods. Besides the new bootstrap method, four other resampling methods were tested. The first one is the bootstrap with replacement proposed by McCarthy and Snowden (1985) for which a correction factor for the finite population is used. The second one is the bootstrap without replacement, which consists of creating an artificial population from the initial sample and drawing bootstrap samples with the same design as the initial one (Gross 1980; Chao and Lo 1985). In the cases of simple random sampling without replacement or unequal inclusion probability sampling design as initial sampling designs, the third method is the rescaled bootstrap of Rao and Wu (1988). For the Poisson sampling design, we used the Patak and Beaumont (2009) method. Nonlinear functions of interest were also tested: the ratio of two totals, the median, and the Gini index. For these functions of interest, the variances under the simulations, say the Monte Carlo variances, were considered as the true variances of the estimators. In the case where the total was the function of interest, the results were directly compared with the variance of the total that can be exactly computed, and not with the Monte Carlo simulation variance. After drawing the bootstrap samples, the estimators, their variances and the means of these variances were computed for each of the initial samples and 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.

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 %

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

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

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

Table 1. Performance of resampling methods in Poisson sampling

Poisson	L	U	ER	Relative bias	RRMSE
Total					
New method	0.5	4.7	5.2	−0.0278	38.5813
Bootstrap WR	10.1	16.2	26.3	−76.4830	78.6988
Bootstrap WOR	4.9	5.4	10.3	−35.4241	36.1937
Method of Patak–Beaumont	1.0	6.1	7.1	−2.8247	40.0502
Median					
New method	3.9	6.2	10.1	0.4701	60.1267
Bootstrap WR	2.3	4.3	6.6	−12.9935	50.2141
Bootstrap WOR	1.9	0.6	2.5	66.7149	113.1575
Method of Patak–Beaumont	3.1	4.8	7.9	8.0193	64.6926
Gini					
New method	1.1	9.8	10.9	−5.3805	38.4937
Bootstrap WR	0.0	5.2	5.2	15.3095	44.8152
Bootstrap WOR	3.5	13.9	17.4	−41.5459	48.1382
Method of Patak–Beaumont	0.6	8.8	9.4	8.1915	65.8452
Ratio					
New method	2.3	4.0	6.3	1.6710	59.6199
Bootstrap WR	0.6	2.6	3.2	−4.8825	49.1502
Bootstrap WOR	8.3	6.2	14.5	−45.2226	48.6318
Method of Patak–Beaumont	1.8	4.8	6.6	8.0236	76.6924

- Total error rate (ER) in %

$$\mathrm{ER} = 100 - \frac{100}{sim}\sum_{i=1}^{sim} I\Big[\widehat{\theta} - 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta}^*)} \le \theta \le \widehat{\theta} + 1.96 \times \sqrt{\mathrm{var}(\widehat{\theta}^*)}\Big],$$

- Relative Bias

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

- Relative Root Mean Squared Error

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

The RB gives a measure of the bias of the estimator of variance. The RRMSE measures its accuracy. 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. Tables 1, 2, and 3 present the numerical performances of the estimators of variance for the three sampling designs, the four functions of interest and the four resampling methods.

Table 1 presents the outcomes achieved using the Poisson sampling design with inclusion probabilities proportional to variable z. The variance estimator provided by the proposed method is unbiased for the total and for the other considered function, it is nearly unbiased according to the MC simulation. The relative bias are small, even for the Gini index (around −5%). For the total and the ratio, the total error rates are about 5%, and for the two other functions of interest about 10%. The bootstrap with replacement is clearly inefficient for the total. In fact, despite the use of a correction factor, the bootstrap with replacement with fixed sample size cannot catch the variance due to the randomness of the sample size of the Poisson sampling design. The variance estimator can thus largely underestimate the true variance. For the other functions of interest, the bootstrap with replacement provides a relatively high coverage rate, but the estimators themselves are biased. With regard to the bootstrap without replacement, the variance estimators are also strongly biased. For the total, the Gini index and the ratio, the variance estimators underestimate the true variance, and give lower coverage rates. For the median, the coverage rate is 97.5% which is only due to the large overestimation of the variance. In general, the performance of the proposed method and the method of Patak and Beaumont (2009) are equivalent. 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%. We can conclude that the new method provides essentially the same results as the others, but its application is simpler: it does not require a correction factor, rescaling or artificial population.

Table 2 shows the results of the applications of resampling methods for simple random sampling without replacement. Here, the original sampling design has a fixed sample size, which explains why the bootstrap with replacement method performs better. Instead of the method of Patak and Beaumont (2009) dedicated to Poisson sampling, we have used the rescaled bootstrap proposed by Rao and Wu (1988). The simulations show that, for the total error rates, the bootstrap with replacement method performs slightly better than the three others, but the coverage rates provided by these others are also between 93% and 94% for each function of interest. The lower and upper error rates for each method and for each function of interest show the same behavior: the distributions are skewed right. There are small biases, positive in the case of the total, the median and the ratio of two totals, except for the rescaled bootstrap method, where the variance of the median is underestimated. For the Gini index, the first three methods give an

Table 2. Performance of resampling methods in simple random sampling without replacement sampling design

SRSWOR	L	U	ER	Relative bias	RRMSE
Total					
New method	1.3	6.3	7.6	5.9195	35.9356
Bootstrap WR	0.0	4.1	4.1	6.5763	36.3808
Bootstrap WOR	1.2	6.3	7.5	4.6716	35.4567
RW Bootstrap	1.0	6.5	7.5	0.6130	33.0132
Median					
New method	1.8	6.4	8.2	9.4256	56.6512
Bootstrap WR	0.5	4.4	4.9	3.8235	49.2184
Bootstrap WOR	2.1	6.1	8.2	10.7279	58.4537
RW Bootstrap	1.9	6.1	8.0	−1.5549	49.6286
Gini					
New method	1.7	5.0	6.7	−4.4216	17.6308
Bootstrap WR	0.6	2.5	3.1	−2.5877	18.1130
Bootstrap WOR	1.7	5.6	7.3	−3.4073	19.4626
RW Bootstrap	0.7	7.8	8.5	12.3624	42.5067
Ratio					
New method	1.7	4.2	5.9	1.2438	28.5170
Bootstrap WR	0.2	2.6	2.8	3.0868	29.1121
Bootstrap WOR	1.7	4.3	6.0	1.3686	28.5030
RW Bootstrap	1.8	4.8	6.6	0.0379	27.1146

estimator that underestimates the true variance, in contrast to the rescaling bootstrap method. In general, for simple random sampling without replacement, there is no crucial difference in performance between the resampling methods. They all provide a slightly biased estimator, with relatively high coverage rates—around 94%—and the variabilities of the variance estimators are also similar.

Table 3 shows the performance of resampling methods under a maximum entropy design with inclusion probabilities proportional to variable z. In the proposed bootstrap method, the second approximation (9) is used, which gives us $\phi_k = \pi_k$. In the case where the function of interest is the total, the new method gives an unbiased estimator with a coverage rate of 92.2%. The bootstrap with replacement method provides a lower error rate and thus a higher coverage rate. However it is due to a larger estimated confidence interval caused by a slight overestimation of the variance. The bootstrap without replacement method with an artificial population and the rescaling bootstrap method strongly underestimate the variance and consequently give a smaller coverage rate. The RRMSE are essentially the same, and again, the distributions of the estimators are skewed right. Concerning the median, the variance estimator of the new method is unbiased while the bootstrap with replacement method, and the rescaled bootstrap method underestimate the variance. The bootstrap without replacement method overestimates it. For the Gini index, the new method and the bootstrap without replacement method perform almost identically: the estimators of the variance are slightly biased (1%–3% in absolute value as relative bias) with a coverage rate of around 92%–93%. The coverage rate provided by the bootstrap with replacement method is larger, but the variance estimator is biased. The rescaled bootstrap method strongly underestimates the variance, which is the reason why the error rate is higher. Concerning the ratio, the estimator under the new resampling method has a small negative bias. In contrast, the bootstrap with replacement method gives an unbiased estimator and the bootstrap without replacement method gives a variance estimator that is 41% larger than the true variance. To summarize these results: the new method performs at least as well as the other methods considered. At the same time, it is simpler and does not require any additional calculation to estimate the variance of the estimators.

Table 3. Performance of the resampling methods in maximum entropy sampling design

UPWOR	L	U	ER	Relative bias	RRMSE
Total					
New method	0.4	7.4	7.8	−0.9515	35.8027
Bootstrap WR	0.0	2.8	2.8	6.8616	34.9417
Bootstrap WOR	3.1	8.8	11.9	−22.6490	33.1929
RW Bootstrap	2.8	10.6	13.4	−36.7334	49.8869
Median					
New method	3.5	6.8	10.3	0.9405	58.6158
Bootstrap WR	1.3	5.0	6.3	−12.9157	48.5572
Bootstrap WOR	0.2	0.0	0.2	233.5629	280.1593
RW Bootstrap	16.6	19.0	35.6	−71.0074	72.6283
Gini					
New method	2.1	5.6	7.7	−3.8518	7.1675
Bootstrap WR	1.0	3.1	4.1	−12.9006	5.5983
Bootstrap WOR	1.3	5.2	6.5	−1.2589	3.3370
RW Bootstrap	7.7	16.7	24.4	−64.8759	65.8130
Ratio					
New method	2.6	3.6	6.2	−2.3080	36.2905
Bootstrap WR	1.2	1.5	2.7	1.0800	30.9940
Bootstrap WOR	2.0	0.7	2.7	41.4054	53.3239
RW Bootstrap	15.3	13.5	28.8	−71.3400	71.9911

These simulations show that the new bootstrap method works at least as well as the usual bootstrap methods. In Poisson sampling design, the inefficiency of the bootstrap with replacement is clear. It is due to the randomness of the sample size. In general, the new method provides an unbiased or a slightly biased estimator with a coverage rate between 89% and 95% for each of the functions of interest under each sampling design considered. Besides having at least the same performance as the other methods, the main advantage of the new method is that it does not require rescaling, correction factors or an artificial population. Thus, the samples can be directly used to compute the variance of the functions of interest.

12. DISCUSSION

The main idea driving the new methodology presented in this article is that if the original sample is drawn with replacement, the one–one sampling design can be directly used in the bootstrap method even if the units are selected with unequal probabilities. If it is drawn without replacement, the variances are smaller than that of a design with replacement and thus a portion of the resampled units are selected without replacement and another is selected according to a one–one design in order to achieve the correct variance. The implementation of selecting resampled units according to a mixture of sampling designs is straightforward and extremely rapid. It consists in computing the sample sizes of the different components of the mixtures,

and then proceeds to select the bootstrap samples, which do not need to be rescaled. The Horvitz–Thompson weights remain unchanged from the original sample.

The simulations show that the classical bootstrap with replacement is not appropriate under unequal probability sampling without replacement, or if the sample size is random. For simple random sampling without replacement, bootstrap with replacement requires a rescaling factor. The class of methods based on the construction of artificial populations has limitations in its time-consuming execution due to its intricacy. In addition, inaccuracy may arise due to rounding problems arising from the multiplication of sample units by the inverse of their inclusion probabilities which are almost never integer (Holmberg 1998). This problem is bypassed in the methodology proposed in the present work. Regarding the method of Rao and Wu (1988), the bootstrap values need not be values from the original sample because of the redefinition technique; although this indeed provides unbiased estimators, difficulties may arise in cases of calibration, reweighting, and imputation.

The method of Patak and Beaumont (2009) entails noninteger weights that may even be negative, which can lead to bootstrap estimations that are not intuitive. This problem is mitigated via a rescaling method, but requires a rescaling factor for the variance, which also presents difficulties under imputation, calibration, and weighting for total nonresponse. The work proposed here can be seen as a variant of the method of Patak and Beaumont (2009), but we impose weights that are positive and integer.

The use of artificial populations produces the correct variance, but, as shown in simulation studies, can be cumbersome and time consuming. The present work avoids these difficulties and attains bootstrap samples in a direct manner that have precisely the same weights as in the original sample, and do not present any of the previous limitations when weighting, calibration or imputation is required.

[Received December 2009. Revised December 2010.]

REFERENCES

Antal, E., and Tillé, Y. (2010), "Simple Random Sampling With Over-Replacement," *Journal of Statistical Planning and Inference*, 141, 597–601. [536]

Berger, Y. G. (1998), "Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling," *Journal of Official Statistics*, 14, 315–323. [535]

Booth, J. G., Butler, R. W., and Hall, P. (1994), "Bootstrap Methods for Finite Populations," *Journal of the American Statistical Association*, 89, 1282–1289. [534]

Brewer, K. R. W. (1975), "A Simple Procedure for πpswor," *Australian Journal of Statistics*, 17, 166–172. [535]

Brewer, K. R. W., and Donadio, M. E. (2003), "The High Entropy Variance of the Horvitz–Thompson Estimator," *Survey Methodology*, 29, 189–196. [535]

Chao, M.-T., and Lo, S.-H. (1985), "A Bootstrap Method for Finite Population," *Sankhyā, Ser. A*, 47, 399–405. [534,540]

Deville, J.-C., and Tillé, Y. (2005), "Variance Approximation Under Balanced Sampling," *Journal of Statistical Planning and Inference*, 128, 569–591. [536]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [534]

Gross, S. T. (1980), "Median Estimation in Sample Surveys," in *Proceedings of the Survey Research Section*, American Statistical Association, pp. 181–184. [534,540]

Hájek, J. (1981), *Sampling From a Finite Population*, New York: Marcel Dekker. [536]

Henderson, T. (2006), "Estimating the Variance of the Horvitz–Thompson Estimator," MSc thesis, School of Finance and Applied Statistics, The Australian National University. [535]

Holmberg, A. (1998), "A Bootstrap Approach to Probability Proportional-to-Size Sampling," in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 378–383. [534,543]

Kuk, A. Y. C. (1989), "Double Bootstrap Estimation of Variance Under Systematic Sampling With Probability Proportional to Size," *Journal of Statistical Computation and Simulation*, 31, 73–82. [534]

Madow, W. G. (1949), "On the Theory of Systematic Sampling, II," *Annals of Mathematical Statistics*, 20, 333–354. [535]

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 (4), 543–570. [536,539]

McCarthy, P. J., and Snowden, C. B. (1985), "The Bootstrap and Finite Population Sampling," technical report, Public Health Service Publication. [534,540]

Patak, Z., and Beaumont, J.-F. (2009), "Generalized Bootstrap for Prices Surveys," in *Proceedings of the 57th Session of the International Statistical Institute*, Durban, South Africa. [540,541,543]

Rao, J. N. K. (1965), "On Two Simple Schemas of Unequal Probability Sampling Without Replacement," *Journal of the Indian Statistical Association*, 3, 173–180.

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inference for Complex Survey Data," *Journal of American Statistical Association*, 83, 231–241. [534,540,541,543]

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), "Some Recent Work on Resampling Methods for Complex Surveys," *Survey Methodology*, 18, 209–217. [534]

Sampford, M. R. (1967), "On Sampling Without Replacement With Unequal Probabilities of Selection," *Biometrika*, 54, 499–513. [535]

Shao, J., and Tu, D. (1995), *The Jacknife and Bootstrap*, New York: Springer-Verlag. [534]

Sitter, R. R. (1992a), "Comparing Three Bootstrap Methods for Survey Data," *Canadian Journal of Statistics*, 20, 135–154. [534]

——— (1992b), "A Resampling Procedure for Complex Survey Data," *Journal of the American Statistical Association*, 87, 755–765. [534]

Tillé, Y. (2006), *Sampling Algorithms*, New York: Springer. [535]