Statistics Surveys
Vol. 10 (2016) 1–52
ISSN: 1935-7516
DOI: 10.1214/16-SS113

A survey of bootstrap methods in finite population sampling*

Zeinab Mashreghi†

Department of Mathematics and Statistics,
University of Winnipeg
Winnipeg, Manitoba R3B 2E9, Canada
e-mail: z.mashreghi@uwinnipeg.ca

David Haziza and Christian Léger

Department of Mathematics and Statistics,
Université de Montréal
Montréal, Québec H3T 1J4, Canada
e-mail: haziza@dms.umontreal.ca; leger@dms.umontreal.ca

Abstract: We review bootstrap methods in the context of survey data where the effect of the sampling design on the variability of estimators has to be taken into account. We present the methods in a unified way by classifying them in three classes: pseudo-population, direct, and survey weights methods. We cover variance estimation and the construction of confidence intervals for stratified simple random sampling as well as some unequal probability sampling designs. We also address the problem of variance estimation in presence of imputation to compensate for item non-response.

MSC 2010 subject classifications: Primary 62D05; secondary 62F40.
Keywords and phrases: Bootstrap, bootstrap weights, confidence intervals, imputation, missing data, multistage designs, pseudo-population approach, survey sampling, unequal probability sampling, variance estimation.

Received December 2014.

Contents

*This is an original survey paper

†The authors wish to thank the referees, the associate editor and the editor for very useful and constructive comments that led to a much improved version of the paper. They also wish to acknowledge the financial support of NSERC and of CANSSI.

1. Introduction

Statistical agencies, such as the Census Bureau and Statistics Canada, provide researchers with access to detailed micro-level data while preserving confidentiality. Each table of data contains ultimate sample units in its rows and the different variables under study in its columns, plus other columns for survey weights. Parameters of interest can be easily estimated based on these values. However, a crucial step is to use the data to estimate some accuracy measures of a given statistic, such as the variance or a confidence interval, something which is not always easy to obtain through analytical methods. For this purpose, many statistical agencies apply bootstrap resampling methods. The data files prepared by these agencies contain further columns with bootstrap survey weights to be used instead of the original survey weights to compute many bootstrap replicates of the statistic. The Monte Carlo variance estimator of the resulting bootstrap statistics is used to estimate the variance under study while the empirical distribution of the bootstrap statistics can be used to construct a confidence interval. Since the bootstrap methods are readily applicable for many estimators, these methods are attractive from a practical point of view.

The bootstrap was first introduced by Efron (1979) in the context of classical statistics where data are independently and identically distributed (i.i.d.) from an unknown distribution. Since survey data are not necessarily i.i.d., many bootstrap resampling methods have been proposed in the context of survey sampling over the past thirty years. These methods are obtained after making some modifications to the classical i.i.d. bootstrap in order to adapt it for survey data.

Some overviews of bootstrap methods in the context of survey sampling have been published, notably parts of Chapter 6 of Shao and Tu (1995), as well as Lahiri (2003) and Shao (2003) in the special issue of *Statistical Science* celebrating the 25$^{\text{th}}$ anniversary of the bootstrap. But these later two papers emphasized small area estimation and imputed data, respectively. In this full study of the various bootstrap methods in the context of survey sampling, we classify the methods in different groups according to their features and we present them in a unified way that shows the similarities and the differences among the methods in a given group. This comprehensive survey should be useful to researchers who need to use or better understand existing bootstrap methods in survey sampling. It provides sufficient details to help researchers apply the methods or develop new ones.

We classify the various bootstrap methods for complete (full response) survey data in three groups. While this classification is partly arbitrary and could have been done differently, it helps in better understanding them. The first one is the class of the pseudo-population bootstrap methods in which a pseudo-population is first created by repeating the units of the original sample and bootstrap samples are then selected from the resulting pseudo-population, e.g. Gross (1980), Booth et al. (1994) and Chauvet (2007). The second one, called the direct bootstrap methods, consists of directly selecting bootstrap samples from the original sample or a rescaled version of it, e.g. Rao and Wu (1988) and Sitter (1992b). In the third group, called the bootstrap weights methods, an appropriate adjustment is made on the original survey weights to obtain a new set of weights called the bootstrap weights, e.g. Rao et al. (1992) and Beaumont and Patak (2012). Users of public data files prepared by agencies such as Statistics Canada, who are usually not familiar with complex statistical methods, can easily use the generated bootstrap weights. They only need to replace the original weights by the resulting bootstrap weights in the estimator of the parameter of interest to define the bootstrap statistics.

The paper is organized as follows. Basic concepts concerning sampling designs, parameter estimation, and estimation of its variance that will be used in the sequel are introduced in Section 2. The jackknife, the balanced repeated replication and the i.i.d. bootstrap resampling methods are briefly discussed in Section 3. A detailed presentation of the three classes of bootstrap methods in the context of survey data is the main topic of Section 4. Variance estimation, the construction of confidence intervals, the application of the methods to multistage designs and software implementation will be discussed in this section. Note that the preceding methods are designed for finite population parameters where the population under study is treated as fixed. The bootstrap methods introduced in Section 5 are applicable when the study variables in the finite population are seen as a realization of a statistical model and the goal is to estimate the variance of the estimator of the parameter of that statistical model.

In practice, we often must be able to deal with imputed data which are used to compensate item non-response. Treating imputed data as true observations may lead to an underestimation of the variance. Therefore, some bootstrap methods that account for the added variability due to item non-response and imputation have been proposed and are studied in Section 6. We conclude in Section 7.

Note that we do not discuss Bayesian bootstrap methodology. Interested readers can read about its application in a survey context in Aitkin (2008) and Carota (2009).

2. Preliminaries

In this section, we introduce the notation as well as the basic concepts of survey sampling. Given that the bootstrap is to be used for estimators more complex than the total or the mean, we also define the median and the GREG estimator of the total. Many of the bootstrap methods are designed so that the bootstrap variance of the estimator of the total exactly matches the usual unbiased estimator of its variance. Hence we pay special attention to the variance of the Horvitz-Thompson estimator of total, as well as the approximation of the variance of functions of totals through Taylor linearization.

Let U be a finite population consisting of N distinct units. Let $y_1, \ldots, y_J$ be J study variables and $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{Ji})^\top$ denote the vector of study variables associated with the i^{th} unit, $i = 1, \ldots, N$. We are interested in estimating a finite population parameter, denoted by θ, which is a function of the N values, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$. A simple but important parameter, in the case where $J = 1$, is the population total of a study variable y defined as $\theta \equiv t = \sum_{i \in U} y_i$. Many parameters encountered in practice can be expressed as a function of population totals:

$$\theta = g(t_1, \ldots, t_J) \quad \text{with } t_j = \sum_{i \in U} y_{ji} \text{ for } j = 1, \ldots, J. \tag{2.1}$$

Special cases of (2.1) include the ratio of two population totals, $\theta = t_1/t_2$, and the finite population distribution function

$$F_N(z) = \frac{1}{N} \sum_{i \in U} I(y_i \leq z), \tag{2.2}$$

where $I(A)$ is the indicator function of the event A taking the value 1 when A occurs and 0 otherwise, and z is a real number. A parameter closely related to the distribution function is the finite population median, which is the value separating the higher half of data from the lower half. More formally, the population median m is defined as

$$m = F_N^{-1}(0.5),$$

where $F_N^{-1}(\cdot)$, the inverse function of $F_N(\cdot)$, is defined as

$$F_N^{-1}(b) = \inf \{ y_i | \; F_N(y_i) \geq b; \; i \in U \}, \tag{2.3}$$

with $0 \leq b \leq 1$.

A sample $s \subseteq U$ of (expected) size n, is randomly selected according to a given sampling design $p(s)$ with first-order inclusion probabilities $\pi_i = Prob(i \in s)$. Common sampling designs include simple random sampling without replacement and stratified simple random sampling, which are both fixed size sampling designs. Fixed size sampling designs are those for which the sample size is fixed prior to sampling. While simple random sampling without replacement is seldom used in practice, stratified simple random sampling is widely applied, especially in business surveys. Under this design, the population U is first divided into L non-overlapping strata $U_1, \ldots, U_L$ with N_h units in the h^{th} stratum,

$h = 1, \ldots, L$. Then, a sample s_h of size n_h is selected from U_h according to simple random sampling without replacement, independently across strata. The first-order inclusion probability of unit i in stratum h is n_h/N_h, $h = 1, \ldots, L$. Except in the case of proportional allocation, stratified simple random sampling is an example of an unequal probability sampling design as units in different strata have different inclusion probabilities.

Another unequal probability sampling design is Poisson sampling, which consists of performing N independent *Bernoulli* trials with probability π_i that unit i is selected in the sample. Unlike simple random sampling without replacement and stratified simple random sampling, Poisson sampling is a random size sampling design.

Estimators of finite population parameters are constructed on the basis of the sample values and, possibly, auxiliary information, which is a set of variables collected for the sample units and for which the corresponding total in the population is known. We start by examining the case of a population total t and consider a general linear estimator of the form

$$\hat{t} = \sum_{i \in s} w_i(s) y_i, \tag{2.4}$$

where $w_i(s)$ is a survey weight associated with the i^{th} unit. The Horvitz-Thompson estimator $\hat{t}_{HT}$ (Horvitz and Thompson, 1952), is an important special case of (2.4) with

$$w_i(s) = w_i = \pi_i^{-1}. \tag{2.5}$$

Suppose that a l-vector of auxiliary variables $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{li})^\top$ is available for all the sample units and that the vector of population totals, $t_{\boldsymbol{x}} = \sum_{i \in U} \boldsymbol{x}_i$, is known. Another linear estimator of t is the so-called Generalized REGression (GREG) estimator (Särndal, 2007), $\hat{t}_G$, given by (2.4) with

$$w_i(s) = \pi_i^{-1} \left\{ 1 + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}HT})^\top \hat{\boldsymbol{T}}^{-1} c_i^{-1} \boldsymbol{x}_i \right\}, \tag{2.6}$$

where $\hat{t}_{\boldsymbol{x}HT} = \sum_{i \in s} \pi_i^{-1} \boldsymbol{x}_i$, $\hat{\boldsymbol{T}} = \sum_{i \in s} \pi_i^{-1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top$ and c_i is a known positive constant attached to unit i. Note that the GREG estimator can also be viewed as a function of estimated totals since it can be expressed as

$$\hat{t}_G = \hat{t}_{HT} + \left(t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}HT} \right)^\top \hat{\boldsymbol{\beta}}, \tag{2.7}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} \pi_i^{-1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in s} \pi_i^{-1} \boldsymbol{x}_i c_i^{-1} y_i.$$

We now turn to the case of parameters that can be expressed as functions of totals, $\theta = g(t_1, \ldots, t_J)$. In this case, we use the plug-in principle that consists of replacing each unknown population total by its corresponding estimator. This leads to the so-called plug-in estimator

$$\hat{\theta} = g(\hat{t}_1, \ldots, \hat{t}_J),$$

where $\hat{t}_j = \sum_{i \in s} w_i(s) y_{ji}$ is a linear estimator of t_j; e.g., the Horvitz-Thompson estimator, for $j = 1, \ldots, J$. For example, the ratio of two totals $\theta = t_1/t_2$ may be estimated by $\hat{\theta} = \hat{t}_{1HT}/\hat{t}_{2HT}$.

Similarly, an estimator of the distribution function (2.2) is given by

$$\tilde{F}_n(z) = \frac{1}{\sum_{i \in s} w_i(s)} \sum_{i \in s} w_i(s) I(y_i \leq z)$$

noting that the population size N in the definition of $F_N(t)$ can be expressed as $N = \sum_{i \in U} 1$. It follows that an estimator of the population median, m, is given by

$$\hat{m} = \tilde{F}_n^{-1}(0.5),$$

where $\tilde{F}_n^{-1}(\cdot)$, the inverse function of $\tilde{F}_n(\cdot)$, is defined as in (2.3).

The above discussion suggests that an estimator of a finite population parameter θ can be viewed as a function of the sample units in s and the survey weights; i.e., $\hat{\theta} = \hat{\theta}\left(s; w_1(s), \ldots, w_n(s)\right)$. This will prove useful when studying the bootstrap weights methods described in Section 4.3.

In this paper, with the exception of Section 5, the properties of estimators (e.g., bias and variance) are studied with respect to the design-based approach. In this approach, the population U is held fixed and the properties of estimators are evaluated with respect to repeated sampling.

The expectation and the variance with respect to the design-based approach are defined as

$$E_p\left(\hat{\theta}\right) = \sum_{s \subset U} \hat{\theta}(s) p(s) \quad \text{and} \quad V_p\left(\hat{\theta}\right) = E_p\left\{\left[\hat{\theta} - E_p\left(\hat{\theta}\right)\right]^2\right\},$$

where the subscript p denotes the sampling design. An estimator is design-unbiased if $E_p\left(\hat{\theta}\right) = \theta$. While the Horvitz-Thompson estimator, $\hat{t}_{HT}$, is design-unbiased for t, the GREG estimator, $\hat{t}_G$, is only asymptotically design-unbiased for t; see, e.g., Isaki and Fuller (1982).

We now turn to the variance of point estimators and variance estimation starting with the case of the Horvitz-Thompson estimator. The design-variance of $\hat{t}_{HT}$ is given by

$$V_p\left(\hat{t}_{HT}\right) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} y_i y_j, \tag{2.8}$$

where

$$\Delta_{ij} = \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j},$$

with $\pi_{ij} = Prob(i \in s \;\&\; j \in s)$ denoting the second-order inclusion probability of units i and j in the sample. The variance (2.8) can be estimated unbiasedly by

$$\hat{V}\left(\hat{t}_{HT}\right) = \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} y_i y_j, \tag{2.9}$$

that is, $E_p\left\{\hat{V}\left(\hat{t}_{HT}\right)\right\} = V_p\left(\hat{t}_{HT}\right)$. For example, under simple random sampling without replacement, (2.9) reduces to the textbook variance estimator of $\hat{t}_{HT}$:

$$\hat{V}\left(\hat{t}_{HT}\right) = N^2(1-f)\frac{s^2}{n}, \tag{2.10}$$

where $f = n/N$ is the sampling fraction and

$$s^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y})^2,$$

with $\bar{y} = n^{-1}\sum_{i\in s} y_i$. For Poisson sampling, noting that $\pi_{ij} = \pi_i\pi_j$ for $i \neq j$, (2.9) reduces to

$$\hat{V}\left(\hat{t}_{HT}\right) = \sum_{i\in s}\frac{1-\pi_i}{\pi_i^2}y_i^2. \tag{2.11}$$

In contrast, the variance of the GREG estimator cannot be obtained in closed form, the latter being a complex function of estimated totals. The same is true for parameters that are expressed as functions of totals such as the ratio of two population totals. To overcome this difficulty, we settle for an approximate expression of the design-variance, which is obtained through a first-order Taylor expansion. Suppose that $\hat{\theta}$ is expressed as a function of estimated totals, $\hat{\theta} = g(\hat{t}_{1HT}, \ldots, \hat{t}_{JHT})$, where $g(\cdot)$ is a differentiable function. Under mild regularity conditions, a first-order Taylor expansion of $\hat{\theta}$ leads to

$$\hat{\theta} - \theta = \sum_{i\in s}\pi_i^{-1}z_i - \sum_{i\in U} z_i + O_p\left(n^{-1}\right), \tag{2.12}$$

where

$$z_i = \sum_{j=1}^{J} y_{ji}\left.\frac{\partial g(\hat{t}_{1HT}, \ldots, \hat{t}_{JHT})}{\partial \hat{t}_{jHT}}\right|_{\hat{t}_{1HT}=t_1,\ldots,\hat{t}_{JHT}=t_J} \tag{2.13}$$

is the so-called linearized variable. For instance, in the case of a ratio, $\theta = t_1/t_2$, the linearized variable is $z_i = (y_{1i} - \theta y_{2i})/t_2$. Ignoring the higher-order terms in (2.12), the design-variance of $\hat{\theta}$ can be approximated by (2.8), where y_i is replaced with z_i. That is, the approximate variance of $\hat{\theta}$ is given by

$$AV_p\left(\hat{\theta}\right) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}z_i z_j. \tag{2.14}$$

As mentioned above, the GREG estimator, $\hat{t}_G$, can also be viewed as a function of estimated totals. In this case, the linearized variable (2.13) reduces to

$$z_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}_U, \tag{2.15}$$

with

$$\boldsymbol{\beta}_U = \left(\sum_{i\in U}\boldsymbol{x}_i c_i^{-1}\boldsymbol{x}_i^\top\right)^{-1}\sum_{i\in U}\boldsymbol{x}_i c_i^{-1} y_i.$$

The approximate variance of $\hat{t}_G$ is thus given by (2.14) with z_i given by (2.15). The approximate variance (2.14) is unknown as the linearized variable z depends on unknown quantities. To estimate (2.14), we start by estimating z by $\hat{z}$. For example, in the case of an estimated ratio, $\hat{\theta} = \hat{t}_{1HT}/\hat{t}_{2HT}$, we have $\hat{z}_i = (y_{1i} - \hat{\theta} y_{2i})/\hat{t}_{2HT}$. An estimator of the approximate variance is obtained from (2.9) by replacing y_i with $\hat{z}_i$, which leads to

$$\hat{V}\left(\hat{\theta}\right) = \sum_{i \in s} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \hat{z}_i \hat{z}_j. \tag{2.16}$$

Under mild regularity conditions (e.g., Deville, 1999), the variance estimator (2.16) is asymptotically unbiased for the approximate variance (2.14).

We briefly discuss asymptotic frameworks needed to have valid linearized variance estimates or confidence intervals. In survey sampling, it is not possible to have the sample size n increase to infinity if the population of size N does not increase at the same time. Moreover, there may be several ways in which the population could grow, depending on the sampling design. For instance, if the population is stratified, the number of strata L could remain fixed, but the size of each stratum N_h increases as would be the case if the strata are provinces or states. Alternatively, the size of the strata could remain relatively fixed, but their number increases as would be the case in a business survey where the classification of businesses would be refined along with an increase in the total number of businesses. In the latter case, the sample size in each stratum would necessarily remain relatively small, sometimes as small as $n_h = 2$, and yet many of the sampling fractions $f_h = n_h/N_h$ could be relatively large. So, in general, one considers an increasing sequence, indexed by a parameter δ, of populations U_δ of size N_δ with a sample of size n_δ for $\delta = 1, 2, \ldots$. In some cases, the units in the population are the results of random variables whereas in others they are a fixed sequence with some conditions that guarantee, for instance, that the variance in the population converges to a constant. Hence, different authors use different asymptotic frameworks. See, for instance, Section 1.3 of Fuller (2009) and Krewski and Rao (1981). But often in survey sampling, authors rely on unspecified regularity conditions...

Both variance estimators (2.9) and (2.16) depend on the second-order inclusion probabilities π_{ij}, which may be difficult to obtain for some unequal probability sampling designs. Moreover, the variance estimator (2.16) obtained through a first-order Taylor expansion requires separate derivations for different functions of estimated totals in order to obtain $\hat{z}$. In this context, resampling methods may prove useful.

3. Bootstrap for independently and identically distributed data and earlier resampling methods

The bootstrap method was first proposed by Efron (1979) in classical statistics where data are i.i.d. from a distribution F. He discovered the bootstrap while trying to better understand the jackknife which was first introduced by Quenouille (1956) to reduce the bias of point estimators. It is based on the idea of combining estimates computed by leaving-out one observation at a time, a form of resampling. Tukey (1958) recognized that by combining these leave-one-out estimators in a certain way, it would be possible to compute estimates of the variance of an estimator. Durbin (1959) was first to apply jackknife variance estimation in the context of finite population sampling while Jones (1974) extended the method to handle stratified sampling. Krewski and Rao (1981), Rao and Wu (1985), Wolter (2007) (pages 174–184), Kovar et al. (1988), Rao et al. (1992) and Chapter 6 of Shao and Tu (1995), among others, studied the properties of jackknife methods in stratified sampling. For multistage cluster sampling, deleting a cluster at a time, a form of delete-a-group jackknife, can be used. Rust (1985) and Kott (1998, 2001) are important references for such methods. For unequal probability sampling without replacement, Campbell and Little (1980), Berger and Skinner (2005), Berger (2007), and Escobar and Berger (2013) introduced generalized jackknife variance estimators. One of the major reasons for the early success of the bootstrap is the fact that unlike the jackknife, it can accommodate functionals which are not as smooth as a mean (or a total), such as quantiles; see Miller (1974) for a review of applications of the jackknife variance estimator in an i.i.d. context. In that context, Shao and Wu (1989) introduced a delete-d jackknife whereby d observations are deleted rather than one, leading to a consistent estimator of variance for quantiles provided that the number of deleted observations d grows at an appropriate rate.

Another ancestor of the bootstrap in a survey context is balanced repeated replication (BRR). McCarthy (1969) first introduced the concept in the specialized case of stratified simple random sampling with replacement with two units selected in each stratum. The idea consists of using half samples in each stratum (modifying the survey weights accordingly) and of considering an appropriately chosen balanced set of such repeated replications of the data. This involves Hadamard matrices; see Appendix A of Wolter (2007). Krewski and Rao (1981), Shao and Wu (1992), and Shao and Rao (1993) studied the properties of such estimators. Extending the method to samples of more than two units per stratum turns out to be a challenge. Gurney and Jewett (1975), Gupta and Nigam (1987), Wu (1991), Sitter (1993), and Rao and Shao (1996, 1999) have all contributed to such extensions. A good review of these other resampling-based methods can be found in Rust and Rao (1996).

We now return to the bootstrap method in the context of an i.i.d. sample from a distribution F as it is important to understand how to generalize it to more complex problems, such as in survey sampling.

Let $Y_1, \cdots, Y_n$ be an i.i.d. sample from the unknown F and let θ be a parameter which is estimated by $\hat{\theta}$ a function of the sample. The bootstrap estimates the distribution of $\hat{\theta}$, computed from an i.i.d. sample from F, by the distribution of $\hat{\theta}^*$, computed from an i.i.d. bootstrap sample $Y_1^*, \ldots, Y_n^*$ from the empirical distribution function $\hat{F}_n$, an estimate of the unknown distribution F. The estimate $\hat{F}_n(z)$ for a real number z is given by

$$\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le z).$$

The bootstrap variance estimate is $V^* = V^*(\hat{\theta}^*|Y_1, \cdots, Y_n)$, the conditional variance of $\hat{\theta}^*$ given $Y_1, \cdots, Y_n$. However, this bootstrap variance estimator is usually not a closed form function of $Y_1, \cdots, Y_n$. In practice, we use a Monte Carlo approximation of V^*. The bootstrap algorithm can be depicted as follows:

1. Generate $Y_1^*, \cdots, Y_n^* \overset{i.i.d.}{\sim} \hat{F}_n$, which is equivalent to drawing a simple random sample $\{Y_1^*, \cdots, Y_n^*\}$ with replacement from $\{Y_1, \cdots, Y_n\}$. Let $\hat{\theta}^*$ be the bootstrap statistic computed on the resulting bootstrap sample.
2. Repeat Step 1 a large number of times, B, to get $\hat{\theta}_1^*, \cdots, \hat{\theta}_B^*$.
3. Estimate $V\left(\hat{\theta}\right)$ by

$$\hat{V}_B^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*\right)^2,$$

   where $\hat{\theta}_{(\cdot)}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^*$.

Conditional on the original sample, when the number of bootstrap samples B goes to infinity, the law of large numbers implies that $\hat{V}_B^*$ converges almost surely to V^*.

A straightforward extension of the bootstrap to survey problems is to apply the above i.i.d. bootstrap algorithm to draw s^*, a simple random sample with replacement (SRSWR) of size n, from the original sample s. For $\hat{\theta} = \hat{t}_{HT}$, the bootstrap variance estimator reduces to

$$V^* = N^2\left(\frac{n-1}{n}\right)\frac{s^2}{n}. \tag{3.1}$$

Even in the case of simple random sampling without replacement, the bootstrap method leads to a biased estimator of the variance as (3.1) fails to account for the finite population correction, $1-f$; see expression (2.10). As a result, the bootstrap variance estimator V^* does not reduce to zero in the case of a census, $s = U$, which is somehow embarrassing; see Lahiri (2003). Of course, in this simple situation, a bias-adjusted variance is easily obtained as $(1-f)[n/(n-1)]V^*$ is consistent and unbiased for the true variance. However, for more complex survey designs, the variance estimator (3.1) is biased and adjusting for the bias may be a complex task unlike in the case of simple random sampling without replacement.

Successful application of the bootstrap in a finite population setting requires appropriate modifications. One approach consists of modifying the bootstrap procedure by taking into account the survey design. Instead of estimating the unknown distribution F and selecting i.i.d. samples from the estimated distribution $\hat{F}_n$, it estimates the unknown finite population U and takes bootstrap samples according to the sampling design. These methods will be presented in

Section 4.1. Alternatively, independent sampling as in the original bootstrap can be used, but to reflect the variability resulting from the sampling design, either the data needs to be rescaled or independent subsamples without replacement must be combined. These methods will be presented in Section 4.2. Finally, in Section 4.3, some bootstrap weights methods will be presented where survey weights rather than the original data set are modified.

4. Design-based bootstrap methods for complete survey data

In this section, we study the complete survey data bootstrap methods introduced so far. To better see the similarities and differences between the various methods, we classify them into three main groups. This classification allows us to present for each group a single algorithm that covers its different methods. There is of course a certain degree of arbitrariness in this classification, but it presents interesting and useful insights. In the first group, a pseudo-population is first created by repeating the elements of the original sample, and bootstrap samples are then selected from the resulting pseudo-population following the original sampling scheme. We refer to these methods as the pseudo-population bootstrap methods. The second one consists of selecting bootstrap samples from the original sample or a rescaled version of it using with replacement sampling, a design that might be different from the original sampling design. We call these methods the direct bootstrap methods. In the third group, called the bootstrap weights methods, instead of resampling observations from the original data set to create a bootstrap sample, the sample remains fixed, but a set of bootstrap survey weights is generated by making rescaling adjustments on the original survey weights. The resulting bootstrap weights, combined with the original data set, are used to compute bootstrap estimators.

It is important to note that many of these methods contain tuning parameters that are set so that the resulting bootstrap expectation and variance in the case of the Horvitz-Thompson estimator of the total coincide with the estimate $\hat{t}_{HT}$, and the usual variance estimator presented in (2.9), respectively. Hence, variance estimation plays a key role in defining the bootstrap methods applied to survey sampling. However, bootstrap confidence intervals are presented at the end of Section 4.1 and in subsequent sections.

4.1. Pseudo-population bootstrap methods

As seen in Section 3, the unknown quantity in the classical i.i.d. model of classical statistics is the distribution F. To perform the bootstrap procedure for this model, F is first estimated by the empirical distribution function $\hat{F}_n$, and then i.i.d. observations from $\hat{F}_n$ are generated. In survey sampling, the unknown is the population U from which the sample is drawn. Therefore, under the pseudo-population bootstrap (PPB) approach, U is estimated by creating a pseudo-population via repeating the original sample using principles from the original sampling design. Then, the bootstrap sample is drawn from the resulting pseudo-population using the original sampling design. By obeying the original scheme to draw the bootstrap sample from the pseudo-population, the finite population correction factors, e.g., the $1-f$ in the case of simple random sample without replacement (SRSWOR), are naturally captured by the bootstrap variance estimator. This important property has persuaded many researchers to widely study this approach.

The pseudo-population bootstrap methods for simple random sample without replacement (or stratified simple random sample) for variance estimation will be presented in the next subsection. Confidence intervals and methods for unequal probability sampling designs will be presented in the two following subsections.

4.1.1. Simple random sampling without replacement: Variance estimation

In this section, we discuss the proposed pseudo-population methods for the case of simple random sample without replacement: Booth et al. (1994), Chao and Lo (1994), Bickel and Freedman (1984), Chao and Lo (1985) and Sitter (1992a). To clarify the application of these bootstrap methods, we illustrate how a pseudo-population is constructed through a simple example. Assume that $N = 1000$ and a simple random sample s of size $n = 100$ is taken without replacement from U. A pseudo-population of size N can be created by repeating the sample s, $N/n = 10$ times. This method was first proposed by Gross (1980). However, in reality, N/n is rarely an integer. In this case, a well-known method to build a pseudo-population of size N was proposed by Booth et al. (1994). In this method, they create a pseudo-population, U^*, by first repeating each unit of the original sample s, $k = \lfloor N/n \rfloor$ times. Then, U^* is completed by taking a simple random sample of size $N-nk$ without replacement from s. For example, assuming that $N = 1000$ and $n = 150$, to construct U^*, each unit in s is first repeated $k = \lfloor 1000/150 \rfloor = 6$ times. Then, U^* is completed by taking a simple random sample of size $N - nk = 100$ without replacement. Note that if N/n is an integer, the pseudo-population U^* created under the method of Booth et al. (1994) is exactly the same as that under the method of Gross (1980).

To construct the pseudo-population, all other pseudo-population methods work similarly to the Booth et al. (1994) method, but different designs are used to complete the pseudo-population. The following algorithm presents a general scheme to create the pseudo-population and to select the bootstrap sample for all existing methods. Elements in ***bold*** in the algorithm need to be specified for each method and will be presented in Table 1.

SRSWOR PPB Algorithm:

1. Repeat each unit in the original sample s, $\boldsymbol{k}$ times to create, U^f, the fixed part of the pseudo-population.
2. Draw $\boldsymbol{U}^{c*}$ from s to complete the pseudo-population, U^*. Therefore, $U^* = U^f \cup U^{c*}$. Let θ^* be the bootstrap analogue of the parameter θ computed on the resulting pseudo-population U^*.
3. Take a simple random sample, s^*, of size $\boldsymbol{n'}$ without replacement from U^*.

4. Compute the bootstrap statistic, $\hat{\theta}^*$, on the bootstrap sample s^*.
5. Repeat Steps 2 to 4 a large number of times, B, to get the sets of bootstrap parameters and bootstrap estimates

$$\theta_1^*, \ldots, \theta_B^*, \quad \text{and} \quad \hat{\theta}_1^*, \ldots, \hat{\theta}_B^*.$$

Later, we will explain how this algorithm is applied to compute bootstrap variance estimators. In Table 1, the number of repetitions k, the design to obtain U^{c*} and the bootstrap sample size n' are presented for all procedures.

Table 1
Existing complete data PPB methods for the case of SRSWOR

Existing methods	k	U^{c*}	n'
Booth et al. (1994)	$\lfloor N/n \rfloor$	SRSWOR from s of size $N-nk$	n
Chao and Lo (1994)		SRSWR from s of size $N-nk$	
Bickel and Freedman (1984)		$\dagger \begin{cases} \emptyset, \text{with pr. } q_{bf}\,[a] \\ s, \text{with pr. } 1-q_{bf} \end{cases}$	
Chao and Lo (1985)		As † with pr. q_{cl} [b]	
Sitter (1992a)	$\left\lfloor \frac{N(n-(1-f))}{n^2} \right\rfloor$	As † with pr. q_s [c]	$n-I\,(U^{c*}=\emptyset)$

[a] $q_{bf} = \left(1 - \frac{N-nk}{n}\right)\left(1 - \frac{N-nk}{N-1}\right)$
[b] $q_{cl} = \frac{G(N)-G(n(k+1))}{G(nk)-G(n(k+1))}$ and $G(t) = \left(1 - \frac{n}{t}\right)\frac{t(n-1)}{(t-1)n}$
[c] $q_s = \frac{\frac{1-f}{n(n-1)} - a_2}{a_1 - a_2}$ with $a_1 = \frac{nk-n+1}{n(n-1)(nk-1)}$ and $a_2 = \frac{k}{n[n(k+1)-1]}$

Note that when N/n is not an integer, for the methods of Booth et al. (1994) and Chao and Lo (1994), the size of the pseudo-population is fixed at N, the original population size, but its (conditional) mean varies with each pseudo-population. On the other hand, for the methods of Bickel and Freedman (1984), Chao and Lo (1985) and Sitter (1992a), there is a randomization between two different pseudo-populations made up of either k or $k+1$ copies of the sample s so that in either case, the (conditional) mean of the pseudo-population is the mean of the sample.

Assuming that $\hat{\theta}$ is an unbiased estimator of θ, $E_p(\hat{\theta}) = \theta$, the variance of $\hat{\theta}$ under the sampling design is reduced to

$$V_p(\hat{\theta}) = E_p[(\hat{\theta} - \theta)^2]. \tag{4.1}$$

To estimate $V_p(\hat{\theta})$ through the SRSWOR PPB Algorithm, there are two random components that have to be considered in this bootstrap procedure: the sampling mechanism applied to complete the pseudo-population and the one for choosing the bootstrap sample, indexed by $u*$ and $p*$, respectively.

Using Step 5 of the SRSWOR PPB Algorithm, a natural Monte Carlo bootstrap estimate of variance would be the sample variance of the bootstrap estimates:

$$\tilde{V}_B^* = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*\right)^2, \tag{4.2}$$

where $\hat{\theta}_{(\cdot)}^* = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^*$. This is an approximation of the (total) bootstrap variance $\tilde{V}^*\left(\hat{\theta}^*\right)$:

$$\begin{aligned}\tilde{V}^*\left(\hat{\theta}^*\right) =& E_{u*p*}\left[\hat{\theta}^* - E_{u*p*}\left(\hat{\theta}^*\right)\right]^2 \\ =& E_{u*}E_{p*}\left\{\left[\hat{\theta}^* - E_{u*}E_{p*}\left(\hat{\theta}^*|U^*\right)\right]^2|U^*\right\}.\end{aligned} \tag{4.3}$$

Conditional on the original sample, $\tilde{V}^*\left(\hat{\theta}^*\right)$ is computed by taking the variability of the bootstrap estimates $\hat{\theta}^*$ around the *fixed* value $E_{u*}E_{p*}\left(\hat{\theta}^*|U^*\right)$, estimated in the algorithm by $\hat{\theta}_{(\cdot)}^*$. It is estimating $E_p[(\hat{\theta}-\theta)^2]$, the *sampling* variability of $\hat{\theta}$ around the (fixed) parameter θ in the population. In some of the bootstrap procedures, conditional on the original sample, the value of the bootstrap parameter, θ^*, is random, i.e., changes with each pseudo-population. In such a case, if the bootstrap estimator $\hat{\theta}^*$ is design-unbiased for its bootstrap parameter θ^*, $\tilde{V}^*\left(\hat{\theta}^*\right)$ does not only reflect the sampling variability of $\hat{\theta}^*$ around θ^*, but also the variability of the bootstrap parameter θ^* around its mean $E_{u*}\left(\theta^*\right)$. So, in the case of methods for which the bootstrap parameter θ^* changes with each pseudo-population, an alternative bootstrap estimate of $V_p(\hat{\theta})$ in (4.1) is obtained by computing the sampling variability of the bootstrap estimates $\hat{\theta}^*$ around θ^*. In other words, we estimate $V_p(\hat{\theta})$ by

$$V^*\left(\hat{\theta}^*\right) = E_{u*}E_{p*}\left[\left(\hat{\theta}^* - \theta^*\right)^2|U^*\right]. \tag{4.4}$$

Using the SRSWOR PPB Algorithm, the Monte Carlo approximation of $V^*\left(\hat{\theta}^*\right)$ is

$$\hat{V}_B^* = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \theta_b^*\right)^2. \tag{4.5}$$

Now, we study both bootstrap variance estimators for all five methods. We start by analyzing the bootstrap procedures of Bickel and Freedman (1984), Chao and Lo (1985) and Sitter (1992a). As was shown in Table 1, each bootstrap method uses a different randomization method to select the pseudo-population. In Bickel and Freedman (1984) and Chao and Lo (1985), the pseudo-population is constructed by randomly repeating the original sample $k = \lfloor N/n \rfloor$ or $\lfloor N/n \rfloor + 1$ times. In Sitter (1992a) the number of repetitions k and the bootstrap sample size are different from those in the other methods. In this method, the randomization is done between two pairs of the number of repetitions k and the bootstrap sample size, i.e. between $(k, n-1)$ and $(k+1, n)$ where $k = \lfloor (N/n)\left[1-(1-f)/n\right] \rfloor$. Consequently, in the case of the mean or the median, the bootstrap parameter θ^* does not change with each pseudo-population and both bootstrap variance estimators $\tilde{V}^*\left(\hat{\theta}^*\right)$ and $V^*\left(\hat{\theta}^*\right)$ of (4.3) and (4.4) are identical.

Consider now the two bootstrap variance estimates in (4.3) and (4.4) for the two other bootstrap procedures of Booth et al. (1994) and Chao and Lo (1994) in the case of the population total, beginning with the former.

For the method of Booth et al. (1994), the bootstrap variance estimator $\tilde{V}^*\left(\hat{\theta}^*\right)$ in (4.3) can be written as

$$\begin{aligned}\tilde{V}^*\left(\hat{t}^*_{HT}\right) =& E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right] + V_{u*}\left[E_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]\\ =& \left[\frac{n-1}{n-f} - \frac{1-f\lfloor N/n\rfloor}{N-1}\left(1-\frac{N-n\lfloor N/n\rfloor}{n}\right)\right]N^2(1-f)\frac{s^2}{n}\\ &+ N\left(1-f\lfloor N/n\rfloor\right)\left(1-\frac{N-n\lfloor N/n\rfloor}{n}\right)s^2,\end{aligned} \tag{4.6}$$

where $\hat{t}^*_{HT} = (N/n)\sum_{i\in s^*} y^*_i$ is the bootstrap Horvitz-Thompson estimator of total computed on s^*. It is straightforward to see that the first term of the bootstrap variance estimator in (4.6) is asymptotically unbiased for $V_p\left(\hat{t}_{HT}\right)$, more precisely,

$$E_p\left\{E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]\right\} - V_p\left(\hat{t}_{HT}\right) = O\left(n^{-1}\right)V_p\left(\hat{t}_{HT}\right). \tag{4.7}$$

Moreover, the ratio of the expectation of each component of (4.6) to $V_p\left(\hat{t}_{HT}\right)$ is

$$\frac{E_p\left\{E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]\right\}}{V_p\left(\hat{t}_{HT}\right)} = O(1) \quad \text{and} \quad \frac{E_p\left\{V_{u*}\left[E_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]\right\}}{V_p\left(\hat{t}_{HT}\right)} = O(f).$$

As a result, the second term in (4.6) produces a bias and implies an overestimation of the variance. This bias can be ignored only when the sampling fraction f is negligible. Note that in the case of a negligible f, even the classical i.i.d. bootstrap method works well asymptotically, so there would be no need to consider more sophisticated resampling procedures.

However, as demonstrated in the following, the bootstrap variance estimator around the bootstrap parameter θ^*, $V^*\left(\hat{t}^*_{HT}\right)$ in (4.4), is asymptotically unbiased for $V_p\left(\hat{t}_{HT}\right)$ regardless of the sampling fraction f. Since $E_{p*}\left(\hat{t}^*_{HT}|U^*\right) = \sum_{i\in U^*} y^*_i = t^*$, we have

$$\begin{aligned}V^*\left(\hat{t}^*_{HT}\right) =& E_{u*}E_{p*}\left[\left(\hat{t}^*_{HT}-t^*\right)^2|U^*\right]\\ =& E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]\\ =& \left[\frac{n-1}{n-f} - \frac{1-f\lfloor N/n\rfloor}{N-1}\left(1-\frac{N-n\lfloor N/n\rfloor}{n}\right)\right]N^2(1-f)\frac{s^2}{n}.\end{aligned} \tag{4.8}$$

Therefore, the bootstrap variance estimator $V^*\left(\hat{t}^*_{HT}\right)$ is equal to the first term of $\tilde{V}^*\left(\hat{t}^*_{HT}\right)$, $E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]$, which represents the average over the different pseudo-populations of the sampling variability of the bootstrap estimator $\hat{t}^*_{HT}$; see (4.6). Consequently, an alternative to formula (4.5) to approximate $V^*\left(\hat{t}^*_{HT}\right)$ without resorting to the explicit computation of the bootstrap parameter θ^* consists of replacing Step 5 of the SRSWOR PPB Algorithm by the following two steps:

5'. Repeat Steps 3 and 4 a large number of times, B, to get the bootstrap estimates $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$, leading to

$$\hat{V}^*_B = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \hat{\theta}^*_{(\cdot)}\right)^2,$$

where $\hat{\theta}^*_{(\cdot)} = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*_b$.

6. Repeat Steps 2 to 5 a large number of times, D, to get $\hat{V}^*_{1B}, \ldots, \hat{V}^*_{DB}$, leading to

$$\hat{V}^* = \frac{1}{D}\sum_{d=1}^{D}\hat{V}^*_{dB},$$

Incidentally, this is the algorithm presented by Chauvet (2007). He argues that since the interest is in estimating the *sampling* variability associated with simple random sampling, the extra variability associated with completing the pseudo-population, $V_{u*}\left[E_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]$, is viewed as a parasitic variance. As a result, the bootstrap estimate of variance should be $E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]$ as computed above.

It should be noted that Booth et al. (1994) were interested in constructing a confidence interval for a function of means and obtained asymptotic results for the distribution of the estimator, which is what is needed to study confidence intervals. Even though they do provide an algorithm for the expected value of the bootstrap estimator, they are silent on estimating the variance of an estimator. On the other hand, they do center the bootstrap statistic $\hat{\theta}^*$ around the bootstrap parameter θ^*, which perhaps suggests that if they had explicitly suggested a variance estimator, it would have been $V^*\left(\hat{\theta}^*\right)$, not $\tilde{V}^*\left(\hat{\theta}^*\right)$. More on the implications of this centering problem for confidence intervals in the next subsection.

Like Booth et al. (1994), Chao and Lo (1994) attempt to create a pseudo-population of size N, the same as the original population size. However, Chao and Lo (1994) take a simple random sample *with replacement* to complete the pseudo-population. They construct their method through first principles, using ideas from the method of moments and maximum likelihood to show that when N/n is an integer, the only natural thing to do is to repeat the original sample k times. When N/n is not an integer, they complete the pseudo-population with a simple random sample *with replacement* from the original sample, but while they argued why it should be completed by observations found in the sample, they do not argue why it should be by simple random sampling *with replacement* as opposed to without replacement.

As in Booth et al. (1994), Chao and Lo (1994) do not explicitly explain how to estimate the variance of an estimator from their method. But in the case of the population total, it leads to

$$E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right] = \left[\frac{n-1}{n-f} - \frac{1-f\lfloor N/n\rfloor}{N-1}\left(1-\frac{1}{n}\right)\right] N^2(1-f)\frac{s^2}{n}$$

and

$$V_{u*}\left[E_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right] = N\left(1-f\lfloor N/n\rfloor\right)\left(1-\frac{1}{n}\right)s^2.$$

Since the first term is asymptotically unbiased for $V_p\left(\hat{t}_{HT}\right)$ and the second term cannot be ignored in the case of a non-negligible f, the bootstrap variance estimator in (4.3), which is $E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right] + V_{u*}\left[E_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]$, may lead to an overestimation of the variance $V_p\left(\hat{t}_{HT}\right)$. However, considering the variability of $\hat{t}^*_{HT}$ around the bootstrap parameter t^*, the bootstrap variance estimator in (4.4), leads to an asymptotically unbiased estimator for $V_p\left(\hat{t}_{HT}\right)$ since

$$V^*\left(\hat{t}^*_{HT}\right) = E_{u*}E_{p*}\left[\left(\hat{t}^*_{HT}-t^*\right)^2|U^*\right] = E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right].$$

Coming back to the methods of Bickel and Freedman (1984), Chao and Lo (1985) and Sitter (1992a), the fact that they involve two pseudo-populations of different sizes raises an interesting question for the estimation of the population total. These three methods are designed to estimate the variance of a function of means. Writing the estimator $\hat{t}_{HT}$ of the population total as $\hat{t}_{HT} = N\bar{y}$, where N is the known population size, the bootstrap statistic is $\hat{t}^*_{HT} = N\bar{y}^*$ and $E_{p*}\left(\hat{t}^*_{HT}|U^*\right) = t^* = N\bar{Y}^* = N\bar{y}$, where $\bar{Y}^*$ is the mean of the pseudo-population. As a result, under these three methods, $E_{u*}E_{p*}\left(\hat{t}^*_{HT}|U^*\right) = N\bar{y} = t^*$, so the bootstrap variance estimators presented in (4.3) and (4.4) are equivalent.

Now let N' be the size of the particular pseudo-population randomly selected by the bootstrap method. If the bootstrap statistic was defined using the usual Horvitz-Thompson estimator on a sample of size n' drawn from a pseudo-population of size N', i.e. $\hat{t}^*_{HT} = (N'/n')\sum_{i\in s^*} y^*_i = N'\bar{y}^*$, (4.3) and (4.4) would no longer be equivalent. If this definition was used, $E_{p*}\left(\hat{t}^*_{HT}|U^*\right) = N'\bar{y}$ which depends on the random pseudo-population size N'.

In Table 2, the ratio of the expectation of $V^*\left(\hat{t}^*_{HT}\right) = \tilde{V}^*\left(\hat{t}^*_{HT}\right)$ to $V_p\left(\hat{t}_{HT}\right)$ is presented for these three methods.

Interestingly, there is quite a bit of confusion in the literature regarding the method of Bickel and Freedman (1984), especially the probability q_{bf} of using $\lfloor N/n\rfloor$ copies of the sample as the pseudo-population; see Table 1. Many authors argue that the probability used in this method can be negative making the procedure potentially infeasible. But q_{bf} cannot be negative and the confusion probably comes from a different probability which was used in the unpublished manuscript Bickel and Freedman (1983).

TABLE 2
The ratio of the expectation of $V^*\left(\hat{t}^*_{HT}\right) = \tilde{V}^*\left(\hat{t}^*_{HT}\right)$ to $V_p\left(\hat{t}_{HT}\right)$ in the case of SRSWOR

Existing methods	$E_p\left[V^*\left(\hat{t}^*_{HT}\right)\right]/V_p\left(\hat{t}_{HT}\right)$
Bickel and Freedman (1984)	$(n-1)/(n-f)$
Chao and Lo (1985)	$\left[q_{cl}\left(\frac{k-1}{nk-1}\right)+(1-q_{cl})\left(\frac{k}{n(k+1)-1}\right)\right]\frac{n-1}{1-f}$ [a]
Sitter (1992a)	1

[a] $k=\lfloor N/n \rfloor$, $q_{cl}=\frac{G(N)-G(n(k+1))}{G(nk)-G(n(k+1))}$ and $G(t)=\left(1-\frac{n}{t}\right)\frac{t(n-1)}{(t-1)n}$

To illustrate the accuracy of the five pseudo-population methods in estimating the variance of $\hat{t}_{HT}$ for some specific cases, the ratio of the expectation of both bootstrap variance estimators, $E_p\left[V^*\left(\hat{t}^*_{HT}\right)\right]$ and $E_p\left[\tilde{V}^*\left(\hat{t}^*_{HT}\right)\right]$, to $V_p\left(\hat{t}_{HT}\right)$, which only depend on the population (N) and sample (n) sizes, are presented in Table 3. Four different scenarios made up of two population sizes $N_1 = 100$ and $N_2 = 10000$ with two sampling fractions $f_1 = 6\%$ and $f_2 = 60\%$ are considered.

TABLE 3
The ratio of the expectation of the bootstrap variance estimators $V^*\left(\hat{t}^*_{HT}\right)$ and $\tilde{V}^*\left(\hat{t}^*_{HT}\right)$ to $V_p\left(\hat{t}_{HT}\right)$ assuming $N_1 = 100$, $N_2 = 10000$, f_1=6% and f_2=60%.

PPB methods for SRSWOR	$E_p\left[V^*\left(\hat{t}^*_{HT}\right)\right]$ / $V_p\left(\hat{t}_{HT}\right)$				$E_p\left[\tilde{V}^*\left(\hat{t}^*_{HT}\right)\right]$ / $V_p\left(\hat{t}_{HT}\right)$			
	f_1=6%		f_2=60%		f_1=6%		f_2=60%	
	N_1	N_2	N_1	N_2	N_1	N_2	N_1	N_2
Booth et al. (1994)	0.842	0.998	0.992	1.0	0.843	0.999	1.192	1.2
Chao and Lo (1994)	0.841	0.998	0.989	1.0	0.843	1.001	1.579	1.6
Bickel and Freedman (1984)	0.842	0.998	0.993	1.0	0.842	0.998	0.993	1.0
Chao and Lo (1985)	0.842	0.998	0.993	1.0	0.842	0.998	0.993	1.0
Sitter (1992a)	1	1	1	1	1	1	1	1

For all methods, the ratio $E_p\left[V^*\left(\hat{t}^*_{HT}\right)\right]/V_p\left(\hat{t}_{HT}\right)$ is close to 1 in all scenarios except when $N_1 = 100$ with $f_1 = 6\%$ where the ratio is about 0.84. Note that in this scenario the sample size, $n = 6$, is very small and the results improve rapidly when the sample size increases. For that same scenario, the other ratio $E_p\left[\tilde{V}^*\left(\hat{t}^*_{HT}\right)\right]/V_p\left(\hat{t}_{HT}\right)$ is also about 0.84. In the case of Sitter (1992a), all ratios are exactly 1 because the probability q_s of Table 1 is constructed so that the bootstrap variance estimator is identical to the usual variance estimator in the case of the population mean (or total).

The contribution of the second term in $\tilde{V}^*\left(\hat{t}^*_{HT}\right)$ to the total variance (the difference between $\tilde{V}^*\left(\hat{t}^*_{HT}\right)$ and $V^*\left(\hat{t}^*_{HT}\right)$) is significant in Booth et al. (1994) and Chao and Lo (1994) when the sampling fraction is large ($f_2 = 60\%$), as suggested by the theory above, while this second term is zero for the other three methods so that $E_p\left[V^*\left(\hat{t}^*_{HT}\right)\right] = E_p\left[\tilde{V}^*\left(\hat{t}^*_{HT}\right)\right]$, as discussed earlier. We also note that completing the pseudo-population using without replacement sampling as in Booth et al. (1994) leads to a much smaller bias than the with replacement sampling of Chao and Lo (1994) when $f_2 = 60\%$, however, the bootstrap variance estimator $\tilde{V}^*\left(\hat{t}^*_{HT}\right)$ in both methods implies an overestimation of the variance.

In light of all of these results, it is clear that for the methods of Booth et al. (1994) and Chao and Lo (1994), the bootstrap estimate of variance should be given by $V^*\left(\hat{\theta}^*\right)$ of (4.4).

All methods for the case of simple random sample without replacement can be easily extended to stratified simple random sample without replacement by applying a resampling method independently within strata.

4.1.2. Simple random sampling without replacement: Confidence intervals

Another important statistical problem is the construction of a confidence interval for a parameter θ based on an estimator $\hat{\theta}$. In survey sampling, the use of the bootstrap for this problem has received less attention than in classical statistics, perhaps because statistical agencies often report coefficients of variation as measures of precision for estimators and this measure requires an estimation of the variance of the estimator.

The construction of *approximate* confidence intervals requires estimates of quantiles of the distribution of $(\hat{\theta} - \theta)/\sqrt{V(\hat{\theta})}$ or of $(\hat{\theta} - \theta)/\sqrt{\hat{V}}$, where $\hat{V}$ is an estimate of $V(\hat{\theta})$, the variance of $\hat{\theta}$, and are based on the inversion of a probability statement of these quantities.

The $1 - \alpha$ level asymptotic confidence intervals are given by

$$[\hat{\theta} - z_{1-\alpha/2}\sqrt{\hat{V}}, \hat{\theta} - z_{\alpha/2}\sqrt{\hat{V}}], \tag{4.9}$$

where z_β is the β-quantile of the standard normal distribution. This interval is based on the approximation of $(\hat{\theta} - \theta)/\sqrt{\hat{V}}$ by a standard normal distribution.

Each estimate $\hat{V}$ of the variance of the estimator leads to a different asymptotic confidence interval. For instance, one could use the linearized variance estimator of (2.14) or the BRR or jackknife estimators briefly discussed in Section 3. But one could also use any of the bootstrap estimate $\hat{V}^*$ introduced so far or that will be introduced in the next subsections.

Using the quantiles of the bootstrap distribution of $(\hat{\theta} - \theta)/\sqrt{\hat{V}}$ instead of those of the standard normal distribution, one obtains bootstrap-t confidence intervals. More precisely, let

$$\hat{K}_{n;*}(x) = Prob^*\left(\frac{\hat{\theta}^* - \theta^*}{\sqrt{\hat{V}^*}} \leq x\right) \tag{4.10}$$

where $\hat{\theta}^*$ and $\sqrt{\hat{V}^*}$ are, respectively, the estimate and its variance estimate computed from the bootstrap observations and θ^* is the value of the parameter under bootstrap sampling. Often, $\theta^* = \hat{\theta}$, the estimate computed from the original sample. But in this section, θ^* is the value of the parameter computed from the particular pseudo-population and so in the case of the methods of Booth et al. (1994) and Chao and Lo (1994), that value changes with each bootstrap sample. The bootstrap-t confidence interval is given by

$$[\hat{\theta} - \sqrt{\hat{V}}\hat{K}_{n;*}^{-1}(1-\alpha/2), \hat{\theta} - \sqrt{\hat{V}}\hat{K}_{n;*}^{-1}(\alpha/2)]. \quad (4.11)$$

Again, there are as many bootstrap-t confidence intervals as there are estimators of the variance $V(\hat{\theta})$. Obviously, if one uses a computer-intensive variance estimator, such as the bootstrap (leading to a double bootstrap) or the jackknife, the confidence interval will require a very large number of computations given that it has to be computed on each bootstrap sample.

A more direct way to use the bootstrap to construct confidence intervals simply uses the bootstrap distribution of $\hat{\theta}^*$. Let $\hat{\theta}^*_{(1)} \le \hat{\theta}^*_{(2)} \le \ldots \le \hat{\theta}^*_{(B)}$ be the ordered bootstrap statistics. The bootstrap percentile confidence interval is (usually) given by

$$[\hat{\theta}^*_{(B\alpha/2)}, \hat{\theta}^*_{(B*\{1-\alpha/2\})}]. \quad (4.12)$$

When the bootstrap parameter θ^* equals the estimate $\hat{\theta}$, this interval is equivalent to

$$[\hat{\theta} + L_{n,*}^{-1}(\alpha/2), \hat{\theta} + L_{n,*}^{-1}(1-\alpha/2)], \quad (4.13)$$

where $L_{n,*}(x) = Prob^*(\hat{\theta}^* - \theta^* \le x)$ is an estimate of $L_n(x) = Prob(\hat{\theta} - \theta \le x)$. Note that in this interval, the left quantile is added to the estimate for the lower bound rather than subtracting the right quantile; see (4.11). This is fine, asymptotically, since the asymptotic distribution of $\hat{\theta}$ is symmetric. But, as with the bootstrap-t intervals, in the case of the bootstrap methods of Booth et al. (1994) and Chao and Lo (1994), the bootstrap parameter θ^* changes with each bootstrap sample and the percentile bootstrap confidence interval would have to be computed according to (4.13) rather than directly through (4.12) as otherwise the confidence intervals will overcover the true parameter θ. As we have previously discussed in the context of variance estimation, the overall dispersion of $\hat{\theta}^*$ as opposed to the dispersion of $\hat{\theta}^* - \theta^*$ is too large; see Table 3. So care must be exercised when applying the percentile bootstrap method.

The discussion of confidence intervals in the papers mentioned so far differs considerably. As discussed previously, Booth et al. (1994) had confidence intervals rather than variance estimation in mind when they introduced their method and they considered bootstrap-t confidence intervals for functions of means where the estimate of the standard error is an explicit function of the sample. Note that $\hat{K}_{n;*}(x)$ is centered at the bootstrap parameter θ^* which changes with each pseudo-population. They do not consider percentile intervals. Sitter (1992a) studies bootstrap-t confidence intervals based on a jackknife variance estimate when his method is applied to nonlinear functions of means while he uses the percentile method for the median. Note that the bootstrap parameter θ^* is fixed for this method. Bickel and Freedman (1984) and Chao and Lo (1985) do not specifically address the construction of confidence intervals, but they both study the asymptotic distribution of a suitably studentized bootstrap version of the estimator $\hat{\theta}$ so that their results would imply that a bootstrap-t confidence interval would work. Finally, Chao and Lo (1994) do not discuss confidence intervals at all.

4.1.3. Unequal probability sampling

We now study two procedures designed for unequal (single-stage) probability sampling designs (UEQPS). The methods of Chauvet (2007) for Poisson sampling and Holmberg (1998) for probability proportional to size sampling, also referred to as πPS, aim to emulate the original sampling design as was the case with simple random sample without replacement. A πPS sample of size n selects the i^{th} element without replacement with probability $\pi_i = nx_i/X$ where $X = \sum_{i=1}^N x_i$ is the total of the size variable x, and so the probability of selection is proportional to the size variable. This sample design is particularly efficient for estimating the total through the Horvitz-Thompson estimator when the correlation between x and y is large. As we have seen in Section 2, in Poisson sampling each element of the population is selected independently in the sample with probability π_i and therefore the sample size is random. We can describe the two methods with this general algorithm. The element in ***bold*** will be specified for each method in the next paragraphs (there shouldn't be any confusion with the usual convention that we have used so far that vectors are represented with bold characters).

UEQPS PPB Algorithm:

1. Repeat the pair (y_i, π_i), $\lfloor \pi_i^{-1} \rfloor$ times for all i in s to create, U^f, the fixed part of the pseudo-population.
2. To complete the pseudo-population, U^*, draw U^{c*} from $\{(y_i, \pi_i)\}_{i \in s}$ using Poisson sampling with inclusion probability $\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$ for the i^{th} pair. Denote the pseudo-population by $U^* = U^f \cup U^{c*} = \{(\check{y}_i, \check{\pi}_i)\}_{i \in U^*}$ where $(\check{y}_i, \check{\pi}_i)$ is the i^{th} pair of the pseudo-population and corresponds to one of the values of the variable obtained from the sample and its corresponding probability of selection according to the sample design.
3. Take the bootstrap sample s^* from U^* using the same sampling design that led to s, but with inclusion probability $\boldsymbol{\pi}_i'$ for the i^{th} unit in U^*, as defined in the sequel.
4. Compute the bootstrap statistic, $\hat{\theta}^*$, on the bootstrap sample s^*.
5. Repeat Steps 3 and 4 a large number of times, B, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$. Let
$$\hat{V}_B^* = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^* \right)^2 ,$$
where $\hat{\theta}_{(\cdot)}^* = B^{-1} \sum_{b=1}^B \hat{\theta}_b^*$.
6. Repeat Steps 2 to 5 a large number of times, D, to get $\hat{V}_{1B}^*, \ldots, \hat{V}_{DB}^*$.

We see that the pseudo-population is constructed the same way for both methods. However, to draw the bootstrap sample, the original sampling mechanism used to draw s from U is applied, but with inclusion probability π_i'. Note that π_i' may be different from the original inclusion probability. The sampling design, the inclusion probability π_i' in Step 3 and the approximation of the bootstrap variance estimator using the resulting Monte Carlo approximations are presented next for both methods.

Chauvet (2007) estimates the variance of the population total $V_p\left(\hat{t}_{HT}\right)$ for Poisson sampling design. To obtain the bootstrap variance estimator of Chauvet, Poisson sampling with the original inclusion probabilities $\pi_i' = \breve{\pi}_i$ in Step 3 of the UEQPS PPB Algorithm is used. Recall that $\breve{\pi}_i$ is the probability of selection of the value $\breve{y}_i$, one of the pairs making the pseudo-population and therefore one of the pairs (y_j, π_j) of the original sample. Under this method, the bootstrap variance estimator is $E_{u*}\left[V_{p*}\left(\hat{\theta}^*|U^*\right)\right]$ which is approximated by

$$\hat{V}^* = \frac{1}{D}\sum_{d=1}^{D} V_{dB}^*.$$

As shown in the previous section, $E_{u*}\left[V_{p*}\left(\hat{\theta}^*|U^*\right)\right] = E_{u*}E_{p*}\left[\left(\hat{\theta}^* - \theta^*\right)^2 |U^*\right]$, where θ^* is the population parameter computed on the resulting pseudo-population in Step 2. Chauvet showed that under Poisson sampling, the proposed bootstrap variance estimator reduces to the usual variance estimator of (2.11) in the case of the total estimator since

$$\begin{aligned}
E_{u*}\left[V_{p*}\left(\hat{t}_{HT}^*|U^*\right)\right] &= E_{u*}\left[V_{p*}\left(\sum_{i\in s^*} \pi_i'^{-1} y_i^* |U^*\right)\right] \\
&= E_{u*}\left(\sum_{i\in U^*} \frac{1-\breve{\pi}_i}{\breve{\pi}_i}\breve{y}_i^2\right) \\
&= \sum_{i\in s} \lfloor \pi_i^{-1} \rfloor \frac{1-\pi_i}{\pi_i} y_i^2 + E_{u*}\left(\sum_{i\in U^{c*}} \frac{1-\breve{\pi}_i}{\breve{\pi}_i}\breve{y}_i^2\right) \\
&= \sum_{i\in s} \lfloor \pi_i^{-1} \rfloor \frac{1-\pi_i}{\pi_i} y_i^2 + \sum_{i\in s}\left(\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor\right)\frac{1-\pi_i}{\pi_i} y_i^2 \\
&= \sum_{i\in s} \frac{1-\pi_i}{\pi_i^2} y_i^2.
\end{aligned}$$

Note that the resulting pseudo-population may not have the same size as the original population size, N. But, letting $\breve{M}_i$ be the number of times unit i appears in U^*, we have $E_p E_{u*}\left(\sum_{i\in s} \breve{M}_i\right) = N$.

Holmberg (1998) proposed his bootstrap method for inclusion probability proportional to size sampling designs, and since the size distribution for the pseudo-population is not the same as the original, the first order inclusion probability used in Step 3 of the UEQPS PPB Algorithm is modified to $\pi_i' = n\breve{\pi}_i / \sum_{j\in U^*} \breve{\pi}_j$. According to the theory done in Holmberg (1998) in the case of Pareto sampling (Rosén, 1997), a special case of a high entropy unequal probability sampling design, it is the total bootstrap variance estimator $\tilde{V}^*(\hat{\theta}^*) = E_{u*}\left[V_{p*}\left(\hat{\theta}^*|U^*\right)\right] + V_{u*}\left[E_{p*}\left(\theta^*|U^*\right)\right]$ which is a good approximation of the variance in the case of the population total.

However, to compute the Monte Carlo variance estimator, he ignores the variability induced by creating the pseudo-population. In the case of Pareto sampling, he takes $D = 1$ in the UEQPS PPB Algorithm, so U_{c*} in Step 2 does not change and the pseudo-population is created once from which a large number of bootstrap samples are taken. As a result, the second term in (4.3) is estimated by zero. In this case, his suggested Monte Carlo approximation of the bootstrap variance estimator is

$$\hat{V}^* = \frac{n}{n-1}\hat{V}_B^*.$$

Chauvet (2007) considered other unequal probability sampling designs. To complete the pseudo-population in Step 2 of the algorithm, he suggests using the sampling design under study with the same probabilities $\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$. In particular, he studied the fixed size rejective sampling (or conditional Poisson sampling). To show that the bootstrap estimate of variance works well in this case, he uses the Hájek approximation for the second order inclusion probability to derive an approximation to the variance of the Horvitz-Thompson estimator of the population total and shows that $E_{u*}\left[V_{p*}\left(\hat{t}^*_{HT}|U^*\right)\right]$ (or $E_{u*}E_{p*}\left[\left(\hat{t}^*_{HT} - t^*\right)^2|U^*\right]$) is asymptotically unbiased for $V_p\left(\hat{t}_{HT}\right)$. The Hájek approximation will be good for rejective sampling as it is a high entropy sampling design. Note that when the original inclusion probabilities are proportional to size, the inclusion probabilities to select the bootstrap sample have to be recalculated on each resulting pseudo-population in the same way that the original inclusion probabilities were computed on U. We conjecture that the method of Chauvet (2007) will perform well for any sampling design belonging to the class of high entropy sampling designs, which includes the Rao-Sampford method (Rao, 1965; Sampford, 1967) and randomized proportional-to-size systematic sampling as special cases.

Interestingly, Chauvet (2007), which is a generalization of Booth et al. (1994), did not recognize that it is not necessary to compute B bootstrap statistics for each of D pseudo-populations to capture the appropriate variance. As we explained in the case of simple random sampling, if one computes the bootstrap parameter θ^* for the pseudo-population, it is possible to obtain the appropriate variance by computing the bootstrap variability of $\hat{\theta}^* - \theta^*$.

While Holmberg (1998) did not address the problem of constructing confidence intervals, Chauvet (2007) computed bootstrap percentile intervals, more specifically percentile intervals constructed from the DB values of $\hat{\theta}_i^*$. As explained in the previous subsection, given that the bootstrap parameter θ^* changes with each pseudo-population, the bootstrap percentile intervals should be computed from the quantiles of $\hat{\theta}_i^* - \theta_i^*$ where the pseudo-population changes with each bootstrap sample; see equation (4.13).

This particular difficulty comes from the random completion of the pseudo-population necessary to obtain exact results for the variance in Poisson sampling. Chauvet (2007) introduced a simplified algorithm where the fixed pseudo-population is made up $[\pi_i^{-1}]$ copies of the pair (y_i, π_i) where $[x]$ is the integer nearest to x and no random completion is done. The rest of the algorithm remains the same. Barbiero and Mecatti (2010) suggested exactly the same thing and called this the "0.5 πPS algorithm". In that paper, they also introduce two other methods to construct a fixed pseudo-population for πPS sampling. As mentioned earlier, the selection probability in such a design is $\pi_i = nx_i/X$ where X is the sum of the x_i. In the creation of the pseudo-population, pairs (y_i, π_i) or equivalently (y_i, x_i) are repeated a number of times according to the first step of the UEQPS PPB Algorithm. Instead of completing the pseudo-population with a random selection from the pairs (y_i, x_i), in these two methods coined "x-balanced", they try to make the sum of the values of x_i in the pseudo-population as close as possible to X, the sum in the population, according to two different (fixed) criteria.

Alternatively, Barbiero et al. (2015) introduced six other bootstrap πPS algorithms based on calibration ideas. Let w_i^* be the number of times that the pair (y_i, x_i) is replicated in the pseudo-population and let $N^* = \sum_{i \in s} w_i^*$ and $X^* = \sum_{i \in s} w_i^* x_i$. In the population, these quantities are N and X. One of their algorithms consists of choosing w_i^* as close to π_i^{-1} so that the two calibration constraints $N^* = N$ and $X^* = X$ are satisfied. The other algorithms consider other calibration (or semi-calibration) constraints.

In addition, Sitter (1992a) extended his method to the Rao-Hartley-Cochran method for probability proportional to size sampling (Rao et al., 1962) in such a way that the bootstrap variance estimate in the linear case is the usual variance estimate. Chao and Lo (1994) also investigated the case of unequal probability sampling design. But their algorithm, based on maximum likelihood ideas where the parameters are the unobserved values in the population, is not really practical and was illustrated on a sample of size 2 from a population of size 5.

4.2. Direct bootstrap methods

The bootstrap methods in this category are based on the idea that the bootstrap samples can be directly drawn from the original data set as in Efron (1979) without requiring the creation of a pseudo-population and mimicking the original sampling design. However, some modifications have to be made so that the bootstrap variability reflects the sampling variability of the original sampling design. Some methods modify the observations while others concatenate independent smaller simple random samples without replacement. First, we focus on the procedures handling the case of simple random sampling without replacement.

The rescaling bootstrap (RSB) method proposed by Rao and Wu (1988) is one of the well-known direct bootstrap methods. In this procedure, a rescaling of the original data set is made before drawing the bootstrap sample leading to a valid estimator of the variance of $\hat{\theta} = g(\hat{t}_{1HT}, \ldots, \hat{t}_{JHT})$, a function of population totals such as a ratio, a correlation coefficient or the generalized regression estimator. Let n' be the bootstrap sample size and $y_i' = \bar{y} + C(y_i - \bar{y})$ be the rescaled y-value for unit i, where

$$C = \sqrt{\frac{n'(1-f)}{n-1}}. \tag{4.14}$$

The bootstrap sample, $s^* = \{y_i^*\}_{i=1}^{n'}$, of size n', is then taken *with replacement* from $s' = \{y_i'\}_{i=1}^{n}$ the set of rescaled data. Then, the bootstrap statistic $\hat{\theta}^* = g(\hat{t}^*_{1HT}, \ldots, \hat{t}^*_{JHT})$ is computed where $\hat{t}^*_{jHT} = (N/n') \sum_{i\in s^*} y^*_{ji}$ for $j = 1, \ldots, J$. To illustrate how this bootstrap method performs for a function of totals, assume that the parameter of interest is the population variance which is a function of two totals:

$$\theta = N^{-1} \sum_{i\in U} y_i^2 - \left(N^{-1} \sum_{i\in U} y_i \right)^2 = N^{-1} t_1 - \left(N^{-1} t_2\right)^2 \tag{4.15}$$

with $(y_{1i}, y_{2i}) = (y_i^2, y_i)$. Therefore, the rescaled values of y_{1i} and y_{2i} are given by $(y'_{1i}, y'_{2i}) = \left(\bar{y}_1 + C\left(y_i^2 - \bar{y}_1\right), \bar{y}_2 + C(y_i - \bar{y}_2)\right)$, where $\bar{y}_1 = n^{-1} \sum_{i\in s} y_i^2$ and $\bar{y}_2 = \bar{y}$. The bootstrap sample is now drawn from $\{(y'_{1i}, y'_{2i})\}_{i=1}^{n}$.

It is worth noting that s^* is drawn with replacement like in Efron (1979), but from a rescaled data set and with a size that may be different from n.

As shown below, the rescaling factor C is chosen so that the variance under resampling matches the usual variance estimator of the population total.

$$\begin{aligned} V_{p*}\left(\hat{t}^*_{HT}\right) &= V_{p*}\left(\frac{N}{n'} \sum_{i\in s^*} y_i^* \right) \\ &= \frac{N^2}{n'} \frac{1}{n} \sum_{i\in s} \left(y_i' - n^{-1} \sum_{j\in s} y_j' \right)^2 \\ &= \frac{N^2 C^2}{n' n} \sum_{i\in s} (y_i - \bar{y})^2 \\ &= N^2 (1-f) \frac{s^2}{n}. \end{aligned}$$

Rao and Wu (1988) showed that an inadequate choice of n' could lead to negative values of $\hat{\theta}^*$ even when $\hat{\theta} \geq 0$ and the parameter of interest is necessarily positive. For example, when the parameter of interest is the population variance given by (4.15), choosing $n' > (n-1)/(1-f)$ might lead to a negative value for $\hat{\theta}^*$. This is because the rescaling is not just done on the y_i, but also separately on the

y_i^2, which after recentering may be negative! However, in this case, by choosing $n' \leq (n-1)/(1-f)$, we have $\hat{\theta}^* \geq 0$.

When applying this method to estimate the variance of the GREG estimator given by (2.7), the auxiliary variables $\boldsymbol{x}$ also need to be rescaled the same way that the study variables are. The bootstrap samples are then selected from the rescaled version of the set of pairs $\{(y_i, \boldsymbol{x}_i)\}_{i \in s}$. The resulting bootstrap variance estimator is asymptotically unbiased for the linearization variance estimator given by (2.14). In addition, Kovar et al. (1988) applied the RSB method to the case of quantiles. To draw bootstrap samples, they use the same rescaled data set that was used for the case of the population mean. The bootstrap estimator of quantile is computed on the resulting bootstrap sample in the same way that it was computed on the original sample.

In the following, a general algorithm for the direct bootstrap methods is presented. In Table 4, the different items from this algorithm noted in ***bold*** are defined for each procedure. To put the various procedures in the same algorithm, we define three quantities. We let C be the rescaling factor of the observations. Also the method of Sitter (1992b), called the mirror-match bootstrap, involves the concatenation of k' simple random samples without replacement of size n''. For the methods involving a single i.i.d. sample of size n', we will use $n''=1$ and $k'=n'$. In other words, setting $n''=1$ in the algorithm described below is equivalent to selecting the bootstrap samples with replacement.

SRSWOR Direct Algorithm:

1. Let $y_i' = \bar{y} + \boldsymbol{C}(y_i - \bar{y})$, for $i = 1, \cdots, n$. Define $s' = \{y_i'\}_{i=1}^n$.
2. Take a simple random sample of size $\boldsymbol{n''}$ without replacement from s'.
3. Repeat Step 2, $\boldsymbol{k'}$ times independently, concatenating all subsamples, to get $s^* = \{y_i^*\}_{i=1}^{n'}$, where $n' = k'n''$.
4. Compute the bootstrap statistic, $\hat{\theta}^* = g(\hat{t}_{1HT}^*, \ldots, \hat{t}_{JHT}^*)$, where $\hat{t}_{jHT}^* = (N/n') \sum_{i \in s^*} y_{ji}^*$ for $j = 1, \ldots, J$.
5. Repeat Steps 2 to 4 a large number of times, B, to get $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
6. Estimate the variance of $\hat{\theta}$ by $V_{p*}\left(\hat{\theta}^*\right)$ or by

$$\hat{V}_B^* = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}_b^* - \hat{\theta}_{(\cdot)}^*\right)^2,$$

where $\hat{\theta}_{(\cdot)}^* = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b^*$.

Table 4 shows that the i.i.d. bootstrap of Efron (1979) overestimates the variance as it fails to capture the without replacement correction factor. McCarthy and Snowden (1985) do the same as Efron (1979), but they recommended a new bootstrap sample size $n' = (n-1)/(1-f)$ to capture the finite population correction factor which yields the customary variance estimator of $\hat{t}$. If the recommended resample size $(n-1)/(1-f)$ is not an integer, they suggest using the closest integer to this value as n'.

TABLE 4
Existing complete data direct bootstrap methods for the case of SRSWOR

Existing methods	C	n''	k'	$\frac{E_p[V_{p*}(\hat{t}^*_{HT})]}{V_p(\hat{t}_{HT})}$
Efron (1979)	1	1	n	$\frac{n-1}{n(1-f)}$
McCarthy and Snowden (1985)	1	1	$\frac{n-1}{1-f}$ [a]	1 [b]
Rao and Wu (1988)	$\sqrt{\frac{k'(1-f)}{n-1}}$	1	Arbitrary [c]	1
Sitter (1992b)	1	$\leq \frac{n}{2-f}$	$\left\lfloor \frac{n(1-f'')}{n''(1-f)} \right\rfloor + I_q$ [d]	1

[a] It may be a non-integer. If so, $n' = \lfloor (n-1)/(1-f) + 0.5 \rfloor$.
[b] Only when $(n-1)/(1-f)$ is an integer
[c] More conditions are required to have a positive $\hat{\theta}^*$ when $\hat{\theta}$ is necessarily positive.
[d] $I_q \sim Bernoulli(q)$ with $q = (\lfloor k \rfloor^{-1} - k^{-1})/(\lfloor k \rfloor^{-1} - \lceil k \rceil^{-1})$, $\lceil k \rceil = \lfloor k \rfloor + 1$, $k = n(1 - f'')/[n''(1-f)]$ and $f'' = n''/n$

As mentioned above, the method of Sitter (1992b) consists of taking a resample without replacement, as in the original sampling scheme, but of size n'' smaller than the original sample size and then repeating this resampling independently $k = n(1-f'')/[n''(1-f)]$ times. The bootstrap sample is obtained by accumulating all these resamples. The number of repetitions k is chosen in such a way that the resulting bootstrap variance matches the usual variance estimate of the population total in (2.9), $V_{p*}\left(\hat{t}^*_{HT}\right) = \hat{V}\left(\hat{t}_{HT}\right)$. Since k is usually not an integer, a randomization between bracketing integers is available as shown in Table 4. Sitter (1992b) showed that this procedure remains valid for the case of a function of totals, but more study is required for more complex parameters such as a population quantile.

Sitter (1992b) also discussed an alternative choice of resample size with $n'' = fn$ such that the resampling fraction $f'' = n''/n$ is the same as the original sampling fraction f. However, this procedure is generally not feasible since both n'' and k are generally not integer values. In this case, two types of randomization between bracketing integers were suggested. In the first one, the bootstrap sample size $n'' = \lfloor fn \rfloor + I_{q'}$ is first fixed, where $I_{q'} \sim Bernoulli(q')$ with $q' = fn - \lfloor fn \rfloor$. Then, a randomization between the integer values of k is done, as presented in Table 4, so that $E(f'') = f$ and $V_{p*}\left(\hat{t}^*_{HT}\right) = \hat{V}\left(\hat{t}_{HT}\right)$. Choosing n'' by this way may lead to $k < 1$. So, this randomization is not valid. In this case, another kind of randomization made between $(\lfloor fn \rfloor, \lfloor k \rfloor)$ and $(\lceil fn \rceil, \lceil k \rceil)$ is presented, where $\lceil \cdot \rceil$ denotes the smallest integer greater than.

All proposed methods can be easily extended to the case of stratified simple random sample without replacement by performing resampling independently within each stratum. Rao and Wu (1988) and Sitter (1992b) also extended their methods to the Rao-Hartley-Cochran method.

Confidence intervals

Consider now the construction of confidence intervals. The methods of Rao and Wu (1988) and Sitter (1992b) both involve a tuning parameter. In the first case, it is the bootstrap sample size n' (with replacement) which has an impact on the rescaling factor of the observations. For Sitter (1992b), it is n'', the size of the without replacement subsample taken from the original sample, which determines the total number of subsamples which will be concatenated to form the bootstrap sample. Whatever value of that parameter is used, these methods have been designed so that the bootstrap estimate of variance is identical to the classical estimate when the estimator is a function of means.

On the other hand, bootstrap confidence intervals will differ depending on the value of the tuning parameter. Consider for instance the RSB method of Rao and Wu (1988) applied to construct a confidence interval for the mean of the distribution. Let $y_{(i)}$ be the i^{th} order statistic of the sample s and let $y'_{(i)}$ of s' be similarly defined. The smallest possible value of the bootstrap statistic $\hat{\theta}^*$ is equal to $y'_{(1)}$ and happens if all n' bootstrap observations are equal to $y'_{(1)}$. But since $y'_{(i)}$ depends on the size of the bootstrap sample n', the distribution of $\hat{\theta}^*$ changes with n' although its variance does not. Consequently, the various bootstrap confidence intervals will vary according to the choice of the tuning parameter, unlike the asymptotic confidence interval based on a bootstrap estimate of variance.

Rao and Wu (1988) have shown that the distribution of the studentized version of the bootstrap statistic, $K_{n;*}(x)$ of (4.10), captures the second-order term of the Edgeworth expansion in the (very) special case of known population strata variances and with replacement sampling, leading to more precise bootstrap-t confidence intervals whenever $n' \approx n-3$. In their paper, they considered bootstrap-t confidence intervals based on a bootstrap estimate of variance (hence a double bootstrap) as well as percentile bootstrap intervals. They also mentioned the possibility of replacing the bootstrap estimate of variance in the bootstrap-t confidence intervals by a jackknife, a BRR or even a linearization variance estimator. Sitter (1992b) also considered percentile and bootstrap-t confidence intervals with his mirror match method.

4.3. Bootstrap weights methods

As discussed in Section 2, an estimator of θ can be viewed as a function of the observations and of the survey weights. Rao et al. (1992) developed the idea of creating bootstrap survey weights rather than drawing the bootstrap sample of observations to compute the bootstrap statistic. In the case of the sample mean, they noted that the bootstrap sample mean $\bar{y}^*$ of the RSB method of Rao and Wu (1988), the mean of the bootstrap observations y_i^*, is a weighted mean of the rescaled observations y'_i where the weights are the number of times that y'_i is in the bootstrap sample. But since y'_i is itself a weighted mean of the original observations y_i, $\bar{y}^*$ is therefore a weighted mean of the original observations y_i. To better understand this statement, let

$$I^*_{ji} = \begin{cases} 1, & \text{if } y^*_j = y'_i = \bar{y} + C(y_i - \bar{y}), \\ 0, & \text{otherwise,} \end{cases} \qquad j = 1, \dots, n'; i = 1, \dots, n.$$

As a result, $\sum_{j \in s^*} I_{ji}^*$ represents the number of times unit i in s is selected in the bootstrap sample under the RSB method. In the case of a population mean, the bootstrap estimator in Rao and Wu (1988) is $n'^{-1} \sum_{j \in s^*} y_i^*$. In the case of simple random sampling without replacement, applying the definition of I_{ji}^*, we have

$$\begin{aligned}
\frac{1}{n'} \sum_{j \in s^*} y_j^* &= \frac{1}{n'} \sum_{j \in s^*} \sum_{i \in s} I_{ji}^* y_i' \\
&= \frac{1}{n'} \sum_{j \in s^*} \sum_{i \in s} I_{ji}^* [\bar{y} + C(y_i - \bar{y})] \\
&= \bar{y} + \frac{C}{n'} \sum_{i \in s} y_i \sum_{j \in s^*} I_{ji}^* - C\bar{y} \\
&= \frac{1}{N} \sum_{i \in s} \left[1 + C \left(\frac{n \sum_{j \in s^*} I_{ji}^*}{n'} - 1 \right) \right] w_i y_i,
\end{aligned}$$

where $w_i = N/n$ is the weight attached to unit i. Therefore, rather than selecting bootstrap observations, Rao et al. (1992) suggested to keep the original observations and create bootstrap weights. This method is attractive to users of public data files prepared by statistical agencies such as Statistics Canada. These agencies provide data sets consisting of columns with the original observations, a column with the original survey weights and B columns of bootstrap weights. As a result, the agencies do not need to provide certain details about the sampling design which could reveal enough information to jeopardize confidentiality.

Bootstrap weights are of the general form

$$w_i^* = a_i^* w_i, \tag{4.16}$$

where a_i^* is computed based on the bootstrap sample. In Rao et al. (1992), the suggested bootstrap adjustments for the case of simple random sampling without replacement are

$$a_i^* = 1 + \sqrt{\frac{n'(1-f)}{n-1}} \left(\frac{n m_i^*}{n'} - 1 \right),$$

where m_i^* is the number of times that the i^{th} element is appearing in the bootstrap sample of size n' selected with replacement from the original sample ($\sum_{i \in s} m_i^* = n'$). Therefore, according to the definition of the random variable $\sum_{j \in s^*} I_{ji}^*$ in the Rao and Wu (1988) method and that of m_i^*, it is clear that the number of times unit i in s is selected in the bootstrap sample has the same distribution in both methods, i.e. $m_i^* \stackrel{D}{=} \sum_{j \in s^*} I_{ji}^*$ where $\stackrel{D}{=}$ indicates equality in distribution. Consequently, we have

$$a_i^* \stackrel{D}{=} 1 + C \left(\frac{n \sum_{j \in s^*} I_{ji}^*}{n'} - 1 \right).$$

TABLE 5
Existing complete data bootstrap weights methods for SRSWOR

Existing methods	Resampling	n'	a_i^*
Rao et al. (1992)	SRSWR	Any [a]	$1+\sqrt{\frac{n'(1-f)}{n-1}}\left(\frac{nm_i^*}{n'}-1\right)$
Chipperfield and Preston (2007)	SRSWOR	$\lfloor n/2 \rfloor$	$1+\sqrt{\frac{\lfloor n/2\rfloor(1-f)}{n-\lfloor n/2\rfloor}}\left(\frac{nm_i^*}{\lfloor n/2\rfloor}-1\right)$
Beaumont and Patak (2012) Bertail and Combris (1997)	–	–	Generate from a distribution with $E^*(\boldsymbol{a}^*)=\mathbf{1}$ and $V^*(\boldsymbol{a}^*-\mathbf{1})(\boldsymbol{a}^*-\mathbf{1})^\top=\boldsymbol{\Sigma}$ [b]
Antal and Tillé (2011)	SRSWOR & one-one	n	m_i^*
Antal and Tillé (2014)	*Bernoulli* & one-one	n	m_i^*

[a] More conditions are required to have positive bootstrap weights.
[b] $\boldsymbol{a}^*=(a_1^*,\dots,a_n^*)$ and $\boldsymbol{\Sigma}=(\Delta_{ij}\pi_i\pi_j/\pi_{ij})$ where $\Delta_{ij}\pi_i\pi_j/\pi_{ij}=-(1-f)/(n-1)$ if $i\neq j$ and $1-f$ if $i=j$.

That is, both methods are equivalent for a function of means (or totals). Note that even if the i^{th} element is not selected, $m_i^*=0$, the associated bootstrap survey weight is nonzero. This is because the rescaled observations y_i' are centered at $\bar{y}$ which involves all observations. If $w_i>0$ for all $i\in s$ and n' is chosen to be less than or equal to $(n-1)/(1-f)$, then the bootstrap weights are all positive.

It is important to understand that the equivalence of the RSB method of Rao and Wu (1988) and of the bootstrap weights version of Rao et al. (1992) only holds for functions of means. In fact, strictly speaking, the RSB method is not even defined when the estimator is not a function of means, but it has for instance been used by Kovar et al. (1988) and Sitter (1992a) for the median where the original data y_k have been rescaled to y_k' and the median was simply computed on the n' selected values of y_k' in the bootstrap sample. It is clear that in this case the RSB bootstrap estimate $\hat{\theta}^*$ is one of the rescaled values y_k' whereas in the bootstrap weights method of Rao et al. (1992), $\hat{\theta}^*$ will be of the original values y_k. So for any other more complex parameter such as a quantile or the Gini index, the two methods differ. And needless to say, the bootstrap is much more needed for such parameters than for simple functions of means.

Letting m_i^* be the number of times that the i^{th} element is appearing in a bootstrap sample selected according to a particular resampling design of size n', Table 5 displays the way a_i^* in (4.16) is computed for different bootstrap weights methods in the case of simple random sample without replacement: Rao et al. (1992), Chipperfield and Preston (2007), Beaumont and Patak (2012) and Antal and Tillé (2011, 2014).

As shown in Table 5, the method of Chipperfield and Preston (2007) introduced a new set of bootstrap weights rescaled on the basis of the number of times that the original units are selected in a simple random sample of size

$n' = \lfloor n/2 \rfloor$ drawn without replacement from s. So, unlike the method of Rao et al. (1992), bootstrap samples are drawn *without replacement*. As a result, $m_i^* = 0$ or 1. Chipperfield and Preston (2007) applied their method and the Rao et al. (1992) method to estimate the variance of GREG estimators. The bootstrap statistics are computed using the following GREG bootstrap weights:

$$w_i^* = a_i^* \pi_i^{-1} \left\{ 1 + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^*)^\top \hat{\boldsymbol{T}}^{*-1} c_i^{-1} \boldsymbol{x}_i \right\},$$

where $\hat{t}_{\boldsymbol{x}}^* = \sum_{i \in s} a_i^* \pi_i^{-1} \boldsymbol{x}_i$ and $\hat{\boldsymbol{T}}^* = \sum_{i \in s} a_i^* \pi_i^{-1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top$. Note that replacing a_i^* by 1 in the expression of w_i^* leads to the usual GREG weights given by (2.6). Both bootstrap variance estimators are asymptotically unbiased to estimate the linear approximation of the variance of total presented in (2.14). Based on empirical results, they showed that the Chipperfield and Preston (2007) method can be significantly more efficient than the bootstrap weights method of Rao et al. (1992) in terms of simulation variance; see also Preston and Chipperfield (2002). As the sample size n increases, Preston and Chipperfield (2002) showed empirically that the difference between both methods vanishes.

A closer look at the Rao et al. (1992) method for the case of SRSWOR reveals that the distribution of $\{m_i^*\}$ is a $Multinomial(n'; \frac{1}{n}, \dots, \frac{1}{n})$, which implies that $E^*(a_i^*) = 1$ and $E^*(a_i^* - 1)(a_j^* - 1) = \Delta_{ij}\pi_i\pi_j/\pi_{ij}$ with $\Delta_{ij}\pi_i\pi_j/\pi_{ij} = 1 - f$ if $i = j$, and $-(1-f)/(n-1)$ otherwise. Therefore, the bootstrap adjustments a_i^* are constructed so that the bootstrap expectation and the bootstrap variance estimator in the case of the estimation of the population total capture the Horvitz-Thompson estimator of total $\hat{t}_{HT}$ and the usual variance estimator $\hat{V}\left(\hat{t}_{HT}\right)$ in (2.9), respectively. Bertail and Combris (1997) and Beaumont and Patak (2012) indicate that if any appropriate distribution is used to generate a_i^* so that

$$E^*(a_i^*) = 1 \quad \text{and} \quad E^*(a_i^* - 1)(a_j^* - 1) = \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}} = \frac{\Delta_{ij}\pi_i\pi_j}{\pi_{ij}}, \qquad (4.17)$$

the first two moments are captured. This type of bootstrap method belongs to the class of the generalized bootstrap method, (e.g., Lo, 1991; Mason and Newton, 1992; Barbe and Bertail, 1995), which was first presented in survey sampling with unequal probability sampling by Bertail and Combris (1997). They suggested generating the vector $\boldsymbol{a}^* = \boldsymbol{1} + \boldsymbol{\Sigma}^{1/2}\tilde{\boldsymbol{a}}^*$ where $\boldsymbol{a}^* = (a_1^*, \dots, a_n^*)$, $\boldsymbol{\Sigma}$ is a $n \times n$ matrix containing $\Delta_{ij}\pi_i\pi_j/\pi_{ij}$ in its i^{th} row and j^{th} column and $\tilde{\boldsymbol{a}}^*$ is a n-vector of independent random variables with mean 0 and variance 1 for all its elements. A simple choice is to generate $\tilde{a}_i^*$ from the standard normal distribution. So, the vector $\boldsymbol{a}^*$ follows a multivariate normal distribution $\mathcal{N}(\boldsymbol{1}, \boldsymbol{\Sigma})$.

In the case of Poisson sampling, the pseudo-population bootstrap method of Chauvet (2007) (see Section 4.1) can be implemented using a bootstrap weights method; see Beaumont and Patak (2012). That is, the creation of a pseudo-population is not required. Rather, bootstrap weights are directly generated from some appropriate distributions so that (4.17) holds. They suggested generating $m_i^* \sim Binomial(\tilde{w}_i, \pi_i)$, where $\tilde{w}_i = \lfloor \pi_i^{-1} \rfloor + I_i^{bp}$ and $I_i^{bp} \sim$

Bernoulli $\left(\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor\right)$, and letting $a_i^* = m_i^*$. The resulting bootstrap parameter and estimator of the population total, θ^* and $\hat{\theta}^*$, from this method and that of Chauvet (2007) have the same distribution; see also Ranalli and Mecatti (2012) for π_i^{-1} integer, for all $i \in s$. Applying the method of Chauvet (2007), we have

$$\begin{aligned}
\hat{\theta}^* - \theta^* &= \sum_{i \in s^*} \pi_i^{'-1} y_i^* - \sum_{i \in U^*} \breve{y}_i \\
&= \sum_{i \in U^*} I(i \in s^*) \breve{\pi}_i^{-1} \breve{y}_i - \sum_{i \in U^*} \breve{y}_i \\
&= \sum_{i \in s} m_i^{'*} w_i y_i - \sum_{i \in s} \left[\lfloor \pi_i^{-1} \rfloor + I(i \in U_{c*}) \right] y_i,
\end{aligned}$$

where $m_i^{'*}$ is the number of times that the i^{th} unit of s is selected in the bootstrap sample from the pseudo-population U^*. Since sample unit i is repeated $\lfloor \pi_i^{-1} \rfloor + I(i \in U_{c*})$ times in U^* and $I(i \in U_{c*}) \overset{D}{=} I_i^{bp}$, it is easy to see that $m_i^{'*} \overset{D}{=} m_i^*$ which confirms that both methods are equivalent in the case of the population total.

In the case of the GREG estimator of total under a given sampling design, assuming that $c_i = \boldsymbol{\lambda}^\top \boldsymbol{x}_i$ in (2.6) with $\boldsymbol{\lambda}$ a vector of known constants so that $c_i > 0$, the GREG survey weights become

$$w_i(s) = \pi_i^{-1} \boldsymbol{x}_i^\top c_i^{-1} \hat{\boldsymbol{T}}^{-1} t_{\boldsymbol{x}}. \tag{4.18}$$

In this case, to compute the corresponding bootstrap statistic, Beaumont and Patak (2012) suggest using their proposed bootstrap adjustments a_i^*, obtained on the basis of the original sampling design, and defining GREG bootstrap weights similar to (4.18) by

$$w_i^* = a_i^* \pi_i^{-1} \boldsymbol{x}_i^\top c_i^{-1} \hat{\boldsymbol{T}}^{*-1} t_{\boldsymbol{x}}, \tag{4.19}$$

where $\hat{\boldsymbol{T}}^* = \sum_{i \in s} a_i^* \pi_i^{-1} \boldsymbol{x}_i c_i^{-1} \boldsymbol{x}_i^\top$. The bootstrap estimator of total is then computed by $\hat{t}^* = \sum_{i \in s} w_i^* y_i$. They showed that the resulting bootstrap variance estimator is approximately equal to the usual variance estimator presented in (2.16). An alternative consists of replacing $t_{\boldsymbol{x}}$ in (4.19) by $\hat{t}_{\boldsymbol{x}HT}$.

In general, some bootstrap adjustments may be negative. To avoid negative bootstrap adjustments a_i^*, Beaumont and Patak (2012) suggested using the following bootstrap adjustments

$$\breve{a}_i^* = \frac{a_i^* + \tau - 1}{\tau},$$

where $\tau \geq 1$ is a small number but large enough so that the scaled bootstrap adjustments are non-negative. Note that $E^*(\breve{a}_i^*) = 1$ and $E^*(\breve{a}_i^* - 1)(\breve{a}_j^* - 1) = \tau^{-2} E^*(a_i^* - 1)(a_j^* - 1)$. Therefore, to have a valid bootstrap estimator for the variance, the resulting bootstrap variance estimator obtained after applying the new bootstrap adjustment $\breve{a}_i^*$ must be multiplied by τ^2. So, this value must be provided to an eventual user.

Antal and Tillé (2011, 2014) have proposed methods applicable for simple random sampling without replacement, Poisson sampling, and unequal probability sampling without replacement.

Confidence intervals

Concerning confidence intervals, Rao et al. (1992) have considered bootstrap-t confidence intervals with a jackknife estimate of variance for functions of means, as well as percentile intervals in the case of the median. Beaumont and Patak (2012) considered percentile intervals in the case of a Poisson design. Bertail and Combris (1997) study asymptotic and bootstrap confidence intervals through a simulation. Unfortunately, they do not explicitly define their intervals, so that it is not clear which one they use, but given the theory that they present, it seems to be the percentile interval. Antal and Tillé (2011) studies an asymptotic interval based on their bootstrap estimate of variance while Antal and Tillé (2014) does not consider the construction of confidence interval at all although the same type of confidence interval could be used. On the other hand, it is not clear that the other types of bootstrap confidence intervals (percentile or t) could be used with either of these methods since they are based on sampling schemes designed to match the variability of the estimator, but not its distribution.

Are all bootstrap methods bootstrap weights methods?

So far, we have organized the various bootstrap methods in three categories: pseudo-population, direct, and bootstrap weights. As argued previously, the rescaling bootstrap method of Rao and Wu (1988) is not a bootstrap weights method unless the estimator can be expressed as a function of means or totals. But all other methods could be expressed in a bootstrap weights approach. There could be two reasons to do this: to have a computationally more efficient way to compute bootstrap weights (than generating the pseudo-population and taking bootstrap samples from it) and to more easily derive the properties of the method. In both cases, the joint distribution of the vector $(a_1^*, a_2^*, \ldots, a_n^*)$ would be needed. But this distribution can be very complicated to write down unless certain special conditions are met. For instance, in the pseudo-population method of Booth et al. (1994) for simple random sampling, if N/n is an integer, the pseudo-population is made up of a fixed number of repetitions of the original sample and bootstrap samples are taken without replacement from it. It can be shown that in this case $(a_1^*, a_2^*, \ldots, a_n^*)$ has a multivariate hypergeometric distribution. But N/n is rarely an integer and in all other cases, the pseudo-population is completed by a (random) sample from the original sample so that it becomes rapidly cumbersome to compute the joint distribution. Hence, from a computational point of view, except when N/n is an integer, it is not clear how one could simulate bootstrap weights from the (complicated) joint distribution of the a_i^* to provide a better implementation of the method.

Moreover, even if we could write it down, it need not give new insight into the variance estimate. For instance, by specifying that the bootstrap resampling is done without replacement from a pseudo-population of size N, one automatically obtains the finite population correction factor $1 - f$ for the bootstrap estimate of variance of the Horvitz-Thompson estimator of the total. This would not be obtained as handily if we worked from a bootstrap weights version of the method.

4.4. Multistage designs

Many surveys are conducted by taking the sample in two or more stages. In such a design, the population U is first partitioned into N_I primary sampling units (PSUs) $U_1, \ldots, U_{N_I}$. In a two-stage design, a sample s_I of PSUs is first selected according to a sampling design $p_I(\cdot)$ and secondary sampling units (SSUs) are further sampled from the i^{th} PSU according to a (possibly different) sampling design $p_i(\cdot|s_I)$. In multistage designs, SSUs are further sampled, and so on. For instance, if the design at the primary level is a census whereas the secondary sampling design is SRSWOR, then the resulting two-stage design is stratified SRSWOR. Alternatively, if the all secondary units from the PSUs selected in the sample s_I are selected in the (ultimate) sample, the resulting design is sometimes called single-stage cluster sample. For more details, see for instance Chapter 4 of Särndal et al. (1997).

The variance of a linear estimator, such as the Horvitz-Thompson estimator of total, involves all second-order probabilities of the primary and secondary designs. In the case of the total, it can be decomposed as the sum of two components representing the variance due to PSU sampling and that due to SSU sampling. If sampling at the first level is with replacement, then the estimate of the total is a weighted combination of the estimate of the total of each PSU in the sample and each of these terms is independent. Hence it is easy to estimate its variance. If sampling is without replacement and the sampling fraction of the primary sampling design is small, using the with replacement estimate of the variance of the total will have a small bias. But if the sampling fraction is large or if the parameter to estimate is more complex, such as a quantile, other estimators of variance must be considered.

Of the methods introduced so far, Sitter (1992a), Sitter (1992b) and Rao and Wu (1988) extended their methods to two-stage sampling (which they refer to as two-stage cluster sampling), while Rao et al. (1992) considered the case of stratified multistage sampling with replacement. In all of these cases, different rescaling factors or tuning parameters were used so that the resulting bootstrap variance estimators match the textbook variance estimator of a linear estimator.

For the case of multistage stratified designs where sampling fractions are large and simple random sample without replacement is used at each stage, a Bernoulli-type bootstrap method was proposed by Funaoka et al. (2006). Under this method, *Bernoulli* trials are applied in each stage of resampling procedure. The bootstrap adjustment for each ultimate unit is the number of times that this unit is selected in the final bootstrap sample. Saigo (2010) extended the methods of Rao et al. (1992), Sitter (1992a) and Sitter (1992b) to stratified three-stage sampling. Drawing the bootstrap samples is of course performed in three stages independently across strata and the rescaling factors used at each stage for the rescaling bootstrap method as well as the number of replications needed at each stage in the mirror-match bootstrap are explicitly presented. They compare these three methods along with that of Funaoka et al. (2006) for variance estimates and confidence intervals of a total and various quantiles through simulations. Preston (2009) extended the bootstrap weights method of Chipperfield and Preston (2007) to multistage stratified designs and compared it to the Bernoulli-type method of Funaoka et al. (2006).

Chaudhuri and Saha (2004), on the other hand, extended the mirror-match method of Sitter (1992b) to two-stage sampling where each stage uses Rao-Hartley-Cochran sampling (Rao et al., 1962).

Chauvet (2007) also generalized his pseudo-population procedure to the case of two-stage sampling design, providing the most general algorithm for such designs. The general idea is simple. For each PSU in the sample, use a pseudo-population method appropriate for the second stage design to recreate a pseudo-PSU by repeating an appropriate number of times the items of the second phase sample from this PSU, for instance using the SRSWOR PPB algorithm if the second stage uses SRSWOR or using the UEQPS PPB Algorithm if it Poisson. After generating pseudo-PSUs for each of the selected PSUs of the first stage sampling, generate an overall pseudo-population by repeating again whole PSUs according to the pseudo-population method appropriate for the first stage design.

Finally, two-phase sampling is different from two-stage sampling in that information on an auxiliary variable is obtained in the (large) sample from the first phase to determine which units will be subsampled to obtain the value of the variable of interest in the second phase. To estimate the population total, a double-expansion estimator and a reweighted expansion estimator are available. In the case of two-phase sampling, Kim et al. (2006) have shown how to use replicate weights computed to reflect variability of the first phase, for instance bootstrap weights, to construct an appropriate replicate-based estimator of variance for these two total estimators.

4.5. Calculation of bootstrap weights: Software implementation

Widespread use of statistical methodology depends on the availability and ease of use of appropriate software. As we have seen, in survey sampling there are many different bootstrap methods and they usually are only valid for a particular type of design. It is therefore important to have enough information to understand which method is used, but this is not always the case. For instance, an important R package in survey sampling is the *survey* package of Lumley (2014) which is the basis of a book: Lumley (2010). This package contains functions to generate replicate weights using the jackknife, BRR or the bootstrap.

Sections 2.3 and 3.2.3 of that book do not specifically mention which bootstrap method is used when they discuss the function *as.svrepdesign*.

The user guide of version 3.30-3 of the package, dated February 20, 2015, presents the functions *bootweights, subbootweights,* and *mrbweights*. According to the documentation, "Bootstrap weights for infinite populations ('with replacement' sampling) are created by sampling with replacement," suggesting that the methods do not take into account that the population is finite. The function *bootweights* is deemed to implement the method of Canty and Davison (1999). But to simplify the discussion, that paper assumes that N/n is an integer and presents what seems like the algorithm of Gross (1980) and the case of a non-integer N/n is not discussed. Since the paper refers to Section 3.7 of Davison and Hinkley (1997), it suggests that *bootweights* implements the method of Booth et al. (1994). The function *subbootweights* seems to implement the Rao et al. (1992) method although the reference is incorrect and no finite population correction is included, i.e., it is as if $f = 0$ in the weights adjustment formula of Table 5 and so is not appropriate if the sampling fraction is large. On the other hand, it is clear that the function *mrbweights* is for the multistage method of Preston (2009). The documentation clearly mentions that "these bootstraps are strictly appropriate only when the first stage of sampling is a simple or stratified random sample of PSUs with or without replacement, and not (eg) for PPS sampling". In fact, Preston's method requires that simple random sampling be used at all stages, not only the first one.

Stata also offers some bootstrap functionalities, including the generation of the Rao et al. (1992) bootstrap weights through the *bsweights* command, see Kolenikov (2010). Of course, this is appropriate for (stratified) simple random sampling only. And as was the case with the *subbootweights* function of R, no finite population correction is available.

So, there are implementations of bootstrap methods for survey sampling in software packages, but it is not always easy to understand what they do exactly, and they often are not appropriate for the problem at hand. Hence more work is needed. Obviously, if bootstrap weights are already available all software packages, including SAS, can be used to calculate bootstrap variance estimates.

5. Bootstrap methods for model parameters

Until now, we have focused on design-based bootstrap methods for finite population parameters. In practice, analysts are often interested in generalizing the conclusions to a universe larger than the finite population under study. For example, one may be interested in studying people's perception of discrimination in their experiences with health care services as a function of characteristics such as race, sex and age. Here, the analyst is not interested in the finite population U currently under study but rather in the process relating these variables. The interest lies in estimating model parameters, also called analytic parameters (e.g., regression coefficients) rather than finite population parameters. An important distinction between finite population parameters and model parameters is that the former may be estimated perfectly provided that a census is conducted and that non-sampling errors such as non-response, measurement errors and coverage errors are absent. In contrast, even with a perfect census, it is not possible to estimate a model parameter perfectly since one faces an infinite population.

In analytic studies, the selected sample can be viewed as the result of a two-stage process: (i) first, the finite population U of size N is generated according to a statistical model, called the superpopulation model, that is, the finite population of size N can be viewed as a realization of the superpopulation model; (ii) then, from the population generated in (i), a sample s is selected according to a given sampling design $p(s)$. Estimators of model parameters are constructed using the sample observations. This begs the question: how to estimate the variance of estimators of model parameters? From the above, it is clear that the variance involves two sources of variability: the first due to the superpopulation model that has generated the finite population U and the second due to the selection of the sample s from U. Application of the bootstrap in this context has been considered in Beaumont and Charest (2012), Wang and Thompson (2012) and Kovacevic et al. (2006). In the sequel, we focus on the method of Beaumont and Charest (2012).

For simplicity, we consider the problem of estimating the regression coefficient $\boldsymbol{\beta}$ in a linear regression model

$$m : y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{x}_i$ is a l-vector of predictors and $\boldsymbol{\beta}$ is a l-vector of unknown parameters. We assume that $E_m(\varepsilon_i) = 0$, $E_m(\varepsilon_i \varepsilon_j) = 0$ if $i \neq j$ and $V_m(\varepsilon_i) = \sigma^2$. Had a census been conducted, an estimator of $\boldsymbol{\beta}$ would be given by

$$\boldsymbol{\beta}_U = \left(\sum_{i \in U} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in U} \boldsymbol{x}_i y_i. \tag{5.1}$$

The estimator (5.1) is often called a census regression coefficient. Since the y-values are only observed for $i \in s$, it is not possible to compute (5.1). An estimator of $\boldsymbol{\beta}_U$ based on the sample units is given by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in s} w_i \boldsymbol{x}_i y_i. \tag{5.2}$$

To derive the variance of $\hat{\boldsymbol{\beta}}$, we first express its total error as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_U \right) + \left(\boldsymbol{\beta}_U - \boldsymbol{\beta} \right).$$

It follows that the total variance of $\hat{\boldsymbol{\beta}}$ is given by

$$V_{mp}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) = E_m V_p \left(\hat{\boldsymbol{\beta}} \right) + V_m \left(\boldsymbol{\beta}_U \right),$$

which involves both the model variability and the sampling variability of $\hat{\boldsymbol{\beta}}$. Under mild regularity conditions, the term $V_p\left(\hat{\boldsymbol{\beta}}\right)$ is of order $O(n^{-1})$, whereas the term $V_m\left(\boldsymbol{\beta}_U\right)$ is of order $O(N^{-1})$; e.g., see Binder (2011). Therefore, the contribution of the term $V_m\left(\boldsymbol{\beta}_U\right)$ to the total variance is negligible if the sampling fraction f is negligible. In this case, the term $V_m\left(\boldsymbol{\beta}_U\right)$ can be omitted and the total variance reduces to

$$V_{mp}\left(\hat{\boldsymbol{\beta}}\right) \approx E_m V_p\left(\hat{\boldsymbol{\beta}}\right). \tag{5.3}$$

In order to estimate $E_m V_p\left(\hat{\boldsymbol{\beta}}\right)$, it suffices to obtain a consistent estimator of $V_p\left(\hat{\boldsymbol{\beta}}\right)$, which represents the sampling variance of a function of totals. To that end, any bootstrap method presented in Section 5, which estimates the sampling variability, can be applied.

We now turn to the case of non-negligible f. For instance, business surveys at Statistics Canada often use stratified sampling where the strata are constructed on the basis of a size variable and while the largest businesses are put in take-all strata, the next largest are in put in take-some strata, often with a large sampling fraction. Using a first-order Taylor expansion, we obtain

$$\begin{aligned} V_{mp}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\simeq \left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1} V_{mp}\left(\sum_{i\in s} w_i \boldsymbol{x}_i e_i\right)\left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1} \\ &= \left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1} E_m V_p\left(\sum_{i\in s} w_i \boldsymbol{x}_i e_i\right)\left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1} \\ &\quad + \left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1} \sum_{i\in U} \boldsymbol{x}_i \boldsymbol{x}_i^\top E_m\left(e_i^2\right)\left\{E_{mp}\left(\hat{\boldsymbol{T}}\right)\right\}^{-1}, \end{aligned} \tag{5.4}$$

where $\hat{\boldsymbol{T}} = \sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$ and $e_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$. In the case of non-negligible f, the last term on the right hand-side of (5.4) is no longer negligible and must be accounted for. A consistent linearization variance estimator of $V_{mp}(\hat{\boldsymbol{\beta}})$ is thus given by

$$\hat{V}\left(\hat{\boldsymbol{\beta}}\right) = \hat{\boldsymbol{T}}^{-1}\hat{V}\left(\sum_{i\in s} w_i \boldsymbol{x}_i e_i\right)\hat{\boldsymbol{T}}^{-1} + \hat{\boldsymbol{T}}^{-1}\left\{\sum_{i\in s} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top e_i^2\right\}\hat{\boldsymbol{T}}^{-1}. \tag{5.5}$$

This begs the question: how to apply the bootstrap method in order to capture both terms in (5.4)? It is clear that applying the bootstrap methods described in Section 4 may lead to an appreciable underestimation of the total variance as the model variability $V_m\left(\boldsymbol{\beta}_U\right)$ is ignored. To overcome this problem, Beaumont and Charest (2012) proposed a bootstrap weights method that accounts for both the sampling and the model variabilities when the sampling design is non informative. Note that a sampling design is non-informative if the distribution of the study variables in the sample is the same as the distribution of these variables in the population, after accounting for $\boldsymbol{x}$. Suppose that the sampling variance in (5.4) is to be estimated through a bootstrap weights method such as the method of Rao et al. (1992). Let $w_i^* = a_i^* w_i$ be the bootstrap weight defined as in Section 4.3 and which addresses the sampling variability. To account for the model variability Beaumont and Charest (2012) suggest making an additional adjustment on the w_i^*. The resulting bootstrap weights are of the form $w_i^{**} = \psi_i^* w_i^* = \psi_i^* a_i^* w_i$, with a_i^* being defined in Section 4.3 and ψ_i^* denoting a random bootstrap adjustment for unit i, whose role is to account for the model variability.

The bootstrap adjustments ψ_i^* are generated independently with expectation equal to 1 and variance equal to

$$V_{o^*}(\psi_i^*) = \sigma_{\psi i}^2 = \frac{w_i}{E_{p*}(w_i^{*2})}, \tag{5.6}$$

where the subscript o^* denotes the distribution of ψ_i^* in the bootstrap samples. To better understand the rationale behind the method of Beaumont and Charest (2012), we first express the bootstrap version of $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_{i \in s} w_i^{**} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)^{-1} \sum_{i \in s} w_i^{**} \boldsymbol{x}_i y_i.$$

Using a first-order Taylor expansion, we obtain

$$V_{p*o*}\left(\hat{\boldsymbol{\beta}}^*\right) \quad \simeq \quad \hat{\boldsymbol{T}}^{-1} V_{p*o*}\left(\sum_{i \in s} w_i^{**} \boldsymbol{x}_i e_i\right) \hat{\boldsymbol{T}}^{-1}, \tag{5.7}$$

where

$$V_{p*o*}\left(\sum_{i \in s} w_i^{**} \boldsymbol{x}_i e_i\right) = V_{p*}\left(\sum_{i \in s} w_i^{*} \boldsymbol{x}_i e_i\right) + E_{p*}\left(\sum_{i \in s} \sigma_{\psi i}^2 w_i^{*2} \boldsymbol{x}_i \boldsymbol{x}_i^\top e_i^2\right).$$

From (5.6), it becomes clear that the total bootstrap variance estimator (5.7) is asymptotically equivalent to the linearization variance estimator (5.5).

To generate ψ_i^*, Beaumont and Charest (2012) suggest using the distribution: $Prob(\psi_i^* = 1 - \sigma_{\psi i}) = 1/2$ and $Prob(\psi_i^* = 1 + \sigma_{\psi i}) = 1/2$. This ensures that ψ_i^* is always non-negative provided that $\sigma_{\psi i} \leq 1$. Note that, in order to compute $\sigma_{\psi i}$, $E_{p*}(w_i^{*2})$ in (5.6) can be easily approximated through a Monte Carlo approximation by taking the mean of the B generated w_i^{*2}.

It is worthwhile to mention that if all the weights w_i are large (implying a small f), $\sigma_{\psi i}^2$ is expected to be small, in which case the contribution of ψ_i^* is expected to be small and, as a result, may be ignored. This conclusion is consistent with the result in (5.3) that the model variability can be ignored if the sampling fraction is small.

We considered the regression model coefficient $\boldsymbol{\beta}$ and its regression estimator (5.2) for ease of presentation. The method can also be used much more generally in the context of survey-weighted estimating equations; see Beaumont and Charest (2012) for details.

We briefly describe the work of the other two papers we mentioned earlier. Many researchers try to understand the hierarchical structure of a population by using multilevel models from complex survey data. Kovacevic et al. (2006) assume that survey weights are available at all sampling levels and that the hierarchy of sampling levels coincides with the hierarchy used in modeling. The proposed methods are modifications of the bootstrap weights method of Rao et al. (1992). Wang and Thompson (2012) focus on weight-inflated estimators of variance components which are often biased. They introduce a pseudo-population based bootstrap method to estimate the bias of the variance components estimators in order to correct them so that they become unbiased with respect to the model and the design, even in the case of an informative design.

6. Bootstrap for missing survey data

Virtually all surveys must face the problem of missing observations due to various reasons. Survey statisticians distinguish unit non-response (when no information is collected on a sample unit) from item non-response (when the absence of information is limited to some variables only). Unit non-response occurs, for example, when the sample unit is not at home or refuses to participate in the survey, while item non-response occurs when the sample unit refuses to respond to sensitive items, may not know the answer to some items, or because of edit failures. In this section, we focus on item non-response, which is typically treated by some form of imputation. In the last two decades, the problem of variance estimation in the presence of imputed data has been widely studied in the literature; see, e.g., Haziza (2009) for a review. It is well known that treating the imputed values as if they were observed values leads to underestimation of the true variance, leading to invalid inferences. In this section, after presenting some useful concepts, some bootstrap methods for imputed survey data will be presented.

6.1. Some useful concepts

Let r_i be the response indicator associated with unit i such that $r_i = 1$ if unit i responds to item y and $r_i = 0$, otherwise. Let

$$y_i^I = r_i y_i + (1 - r_i)\tilde{y}_i,$$

where $\tilde{y}_i$ denotes the imputed value used to replace the missing y_i. Let θ be a finite population parameter, $\hat{\theta}$ be the complete data estimator of θ and $\hat{\theta}^I$ be the imputed estimator obtained after imputation. The imputed estimator $\hat{\theta}^I$ can be computed the same way as the complete data estimator $\hat{\theta}$ using y^I-values instead of the y-values. For example, in the case of a total t, an imputed estimator is

$$\hat{t}^I = \sum_{i \in s} w_i y_i^I.$$

In practice, various imputation methods are used. We distinguish between two classes of imputation methods: the deterministic methods, which are those that yield the same imputed values if the imputation process is repeated, and the random methods that may yield different imputed values if the imputation is repeated. A random method can be viewed as a deterministic method with an added random noise. Most imputation methods encountered in practice are motivated by the general model

$$m : y_i = f(\boldsymbol{x}_i; \boldsymbol{\beta}) + \varepsilon_i, \tag{6.1}$$

where $f(\cdot)$ is a given function, $\boldsymbol{x}$ is a vector of auxiliary variables recorded for all sample units (respondents and non-respondents) and $\boldsymbol{\beta}$ is a vector of unknown parameters. The errors ε_i satisfy

$$E_m(\varepsilon_i) = 0,\ V_m(\varepsilon_i) = \sigma^2 c_i \text{ and } \ cov_m(\varepsilon_i, \varepsilon_j) = 0,\ \forall i \neq j,$$

where σ^2 is an unknown parameter and c_i is a fixed positive constant. For example, deterministic linear regression imputation is motivated by (6.1) with $f(\boldsymbol{x}_i; \boldsymbol{\beta}) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. In this case, the imputed value $\tilde{y}_i$ is given by

$$\tilde{y}_i = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_r, \tag{6.2}$$

where

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{i \in s} w_i r_i \boldsymbol{x}_i {c_i}^{-1} \boldsymbol{x}_i^\top \right)^{-1} \sum_{i \in s} w_i r_i \boldsymbol{x}_i {c_i}^{-1} y_i$$

is the weighted least square estimator of $\boldsymbol{\beta}$ based on the responding units. Mean imputation, whereby the missing values are replaced by the mean of the respondents, $\bar{y}_r = \sum_{i \in s} w_i r_i y_i / \sum_{i \in s} w_i r_i$, is a special case of (6.2) with $\boldsymbol{x}_i = c_i = 1$ for all i.

A frequently used random method is random hot-deck imputation, which consists of imputing a missing value by the value of a respondent selected at random from the set of responding units. More specifically, the imputed values under random hot-deck imputation are

$$\tilde{y}_i = \bar{y}_r + \tilde{\varepsilon}_i, \tag{6.3}$$

where $\tilde{\varepsilon}_i$ takes a value in $\{e_1, \ldots, e_{n_r}\}$ such that $Prob(\tilde{\varepsilon}_i = e_j) = r_j w_j / \sum_{l \in s} r_l w_l$ with $e_j = y_j - \bar{y}_r$ and n_r denoting the number of respondents to item y.

In this section, we assume that the data are Missing At Random (MAR); (Rubin, 1976). The data are MAR if the probability of response to item y is independent of the error term in (6.1) after accounting for the vector of auxiliary variables $\boldsymbol{x}$.

There exist two theoretical frameworks for variance estimation: the customary two-phase framework and the reverse framework. In the two-phase framework, non-response is viewed as a second phase of selection. In the reverse framework, proposed by Fay (1991) and Shao and Steel (1999), the order of sampling and response is reversed. First, the population is randomly divided into a population of respondents and a population of non-respondents according to the non-response mechanism. Then, a random sample is selected from the population (containing respondents and non-respondents) according to the sampling design $p(s)$. Unlike the two-phase framework, the reverse framework requires the additional assumption that the non-response mechanism does not depend on which sample is selected. The reverse framework is particularly useful in the context of bootstrap variance estimation in the presence of imputed data, as we argue in the next section.

6.2. Bootstrap methods for negligible sampling fraction

In this section, we focus on the case of negligible f. In this context, Shao and Sitter (1996) proposed a bootstrap method for handling imputed data. The rationale behind their method is to first select, using any complete data bootstrap method, a bootstrap sample of pairs composed of the original or rescaled imputed data and their corresponding original response status. The bootstrap data with a missing status are then reimputed using the same imputation method that was used in the original sample. To illustrate the Shao-Sitter method, we consider the case of simple random sampling without replacement with the RSB method of Rao and Wu (1988) and mean imputation to compensate for the missing values. The algorithm proceeds as follows:

Shao-Sitter Algorithm:

1. Let n' be the bootstrap sample size and $y_i' = \bar{y}^I + C(y_i^I - \bar{y}^I)$, for all i in s, where $\bar{y}^I = n^{-1}\sum_{i\in s} y_i^I$ and
$$C = \sqrt{\frac{n'(1-f)}{n-1}}.$$
2. Draw a bootstrap sample of pairs $s^* = \{(y_i^*, r_i^*)\}_{i=1}^{n'}$ of size n' with replacement from $\{(y_i', r_i)\}_{i=1}^{n}$.
3. Reimpute the missing values in the bootstrap sample s^* using the respondents in this sample, i.e. define y_i^{*I} as follows
$$y_i^{*I} = \begin{cases} y_i^*, & \text{if } r_i^* = 1, \\ \bar{y}_r^*, & \text{if } r_i^* = 0, \end{cases} \qquad \text{where } \bar{y}_r^* = \frac{\sum_{i\in s^*} r_i^* y_i^*}{\sum_{i\in s^*} r_i^*}, \text{ for } i \in s^*.$$
Let $\hat{\theta}^{*I}$ be the bootstrap statistic based on the observed and imputed bootstrap data.
4. Repeat Steps 2 and 3 a large number of times, B, to get $\hat{\theta}_1^{*I}, \cdots, \hat{\theta}_B^{*I}$.
5. Estimate $V\left(\hat{\theta}^I\right)$ with $V_{p*}\left(\hat{\theta}^{*I}\right)$ or its Monte Carlo approximation $\hat{V}_B^* = (B-1)^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^{*I} - \hat{\theta}_{(\cdot)}^{*I}\right)^2$, where $\hat{\theta}_{(\cdot)}^{*I} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b^{*I}$.

Note that for imputation methods using auxiliary information (e.g., regression imputation), the vector of auxiliary variables $\boldsymbol{x}_i$ also accompanies the pairs (y_i, r_i) in the bootstrap sample and needs to be rescaled similarly to y_i.

In the case of the population total, the bootstrap total estimator is $\hat{t}^{*I} = (N/n') \sum_{i=1}^{n'} y_i^{*I} = N\bar{y}_r^*$. Using a first order Taylor linearization, when the non-response mechanism is uniform, i.e. the response probability $p_i = Prob(r_i = 1) = p_0$ for all $i \in s$, the bootstrap variance estimator $V_{p*}\left(\hat{t}^{*I}\right)$ is approximated by

$$\begin{aligned} V_{p*}\left(\hat{t}^{*I}\right) &\approx V_{p*}\left\{\frac{N}{\hat{p}_0 n'}\sum_{i=1}^{n}\left(m_i^* - \frac{n'}{n}\right)(y_i' - \bar{y}_r)\right\} \\ &= \frac{N^2}{\hat{p}_0^2}\frac{C^2}{n'n}\sum_{i\in s} r_i(y_i - \bar{y}_r)^2 \\ &= N^2\left(\frac{1-f}{\hat{p}_0}\right)\frac{n_r - 1}{\hat{p}_0(n-1)}\frac{s_r^2}{n}, \end{aligned} \tag{6.4}$$

where m_i^* is the number of times that the i^{th} unit in s is selected in the bootstrap sample, $\hat{p}_0 = n_r/n$, the response rate, is the estimator of p_0 and $s_r^2 = (n_r - 1)^{-1}\sum_{i\in s} r_i(y_i - \bar{y}_r)^2$.

At this point, one may be wondering what quantity (6.4) is really estimating. To answer this question, one has to rely on the reverse framework for variance estimation mentioned above. The reverse framework can be used to express the variance of $\hat{\theta}^I$ as the sum of two terms in the case of deterministic imputation. In this case, the total variance of $\hat{\theta}^I$ under deterministic imputation is given by

$$V^{NR}\left(\hat{\theta}^I\right) = EV_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right) + VE_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right), \tag{6.5}$$

where $\boldsymbol{y} = (y_1, \ldots, y_N)^\top$ and $\boldsymbol{r} = (r_1, \ldots, r_N)^\top$ is the vector of response indicators. Under mild regularity conditions, the contribution of the second component to the total variance in (6.5), $VE_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right)/V^{NR}\left(\hat{\theta}^I\right)$, is of order $O(f)$, which is negligible when the sampling fraction, f, is negligible. Therefore, when f is negligible, this component can be omitted from the calculations and only the first component $EV_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right)$ remains to be estimated. To that end, it suffices to estimate $V_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right)$ in an (approximately) unbiased fashion.

Suppose that we are interested in estimating a population total t. Noting that the imputed estimator $\hat{t}^I$ can be expressed as a function of totals, estimating $V_p\left(\hat{t}^I|\boldsymbol{y},\boldsymbol{r}\right)$ reduces to the classical problem of estimating the sampling variance of a function of totals. To that end any complete data variance estimation methods can be used, including Taylor expansion procedures and resampling methods. The bootstrap variance estimator (6.4) is an estimator of $V_p\left(\hat{t}^I|\boldsymbol{y},\boldsymbol{r}\right)$ as the Shao-Sitter method simulates the effect of sampling conditionally on the vector of response indicators $\boldsymbol{r}$ and the bootstrap method reflects this sampling variability. This can be explained by the fact that non-response is not generated in each bootstrap sample before the imputation process is performed; see

Mashreghi et al. (2014). As a result, the bootstrap variance estimator (6.4) can be used if the sampling fraction f is negligible. Also, it is worth noting that (6.4) is approximately unbiased for $V_p\left(\hat{t}^I|\boldsymbol{y},\boldsymbol{r}\right)$ regardless of the validity of the underlying imputation model.

The problem of bootstrap variance estimation in the case of quantiles is discussed in Shao and Chen (1998). The method of Shao-Sitter may lead to a biased estimator in the case of very small stratum sizes. To overcome the problem, Saigo et al. (2001) proposed a modification of the method of Shao and Sitter (1996). Instead of using any complete data bootstrap method like Shao and Sitter (1996), they proposed a new sampling design, called the repeated half-sample bootstrap.

6.3. Bootstrap methods for non-negligible sampling fraction

When the sampling fraction is appreciable, the Shao-Sitter method may lead to a significant underestimation of the variance as the term $VE_p\left(\hat{\theta}^I|\boldsymbol{y},\boldsymbol{r}\right)$ in (6.5) is not accounted for. To overcome this problem Mashreghi et al. (2014) proposed a method called the independent bootstrap in the special case of stratified simple random sample without replacement with uniform non-response in each stratum. Their method consists of selecting bootstrap samples according to a direct bootstrap method (see Section 4.2) and then regenerating non-response within each bootstrap sample, mimicking the initial non-response mechanism, i.e., independent *Bernoulli* trials with the observed response rate. Afterwards, the non-respondents in the bootstrap sample are reimputed using the same imputation method that was used on the original data. Since direct bootstrap methods involve some constants, e.g., C and k' in Table 4, Mashreghi et al. (2014) showed how to modify these constants to obtain an approximately unbiased estimator of the total variance. The modified constants explicitly depend on the response rate as well as the imputation method. For example, in the case of mean imputation with uniform non-response mechanism, the rescaling factor in the method of Rao and Wu (1988) presented in (4.14) has to be replaced by

$$C^I = \sqrt{\frac{n'[1-(n_r/N)]}{n_r - 1}}.$$

Comparing C^I with C in (4.14), we see that n in C is replaced by n_r in C^I, i.e. the number of respondents is used instead of the sample size as the information contained in the sample only comes from the observed values. In this case, the following algorithm leads to the creation of samples of bootstrap imputed data:

1. Let n' be the bootstrap sample size and $y_i' = \bar{y}^I + C^I(y_i^I - \bar{y}^I)$, for all i in s.
2. Draw a bootstrap sample $\{y_i^*\}_{i=1}^{n'}$ of size n' with replacement from $\{y_i'\}_{i=1}^{n}$.
3. Generate $\{r_i^*\}_{i=1}^{n'} \overset{i.i.d.}{\sim} Bernoulli(\hat{p}_0)$, the bootstrap sample of response indicators. Let $s^* = \{(y_i^*, r_i^*)\}_{i=1}^{n'}$.

4. Identify the missing and observed bootstrap data using the regenerated r_i^* and reimpute the bootstrap missing values using the bootstrap respondents and the same imputation method that was used to impute the original sample. Let $\hat{\theta}^{*I}$ be the bootstrap statistic based on the bootstrap imputed data.

Unlike the Shao-Sitter algorithm presented in this section, this algorithm includes an additional step in order to generate non-response within each bootstrap sample. Note that the constant C^I used to rescale the data depends on the imputation method. Mashreghi et al. (2014) present the appropriate constants for several combinations of bootstrap and imputation methods.

7. Conclusion

Efron (1979) revolutionized applied statistics when he introduced the bootstrap methodology to estimate the variance of estimators and construct confidence intervals by resampling the observations of samples of i.i.d. observations. It took close to a decade to apply the bootstrap to more complicated models, such as time series or spatial data. The application to survey methods data also took some time and many approaches and principles were used to define bootstrap methods. The introduction of bootstrap survey weights, Rao et al. (1992), ultimately democratized the use of survey data when statistical agencies such as Statistics Canada prepared public survey data files with columns of bootstrap survey weights that subject-area researchers could use to compute variance estimates for estimators designed to answer research questions of interest.

In this paper, we present the various bootstrap methods that have been introduced for survey methods data by classifying them in three groups. This classification provides interesting and useful insights and allows us to present a single algorithm for all methods of a given group. Variance estimation has been central in the development of many of these methods since they are designed so that the bootstrap variance estimate for the estimate of total is the usual estimate of variance. This is partly due to the fact that many statistical agencies rely on coefficients of variation, which involve the variance of an estimator, to quantify its precision. But the construction of confidence intervals is also an important survey sampling inference problem addressed by the bootstrap.

Many of the pseudo-population bootstrap methods have the advantage that they use the sampling design that generated the original data and so the bootstrap estimator inherits many of its properties. One disadvantage is that they usually require the explicit generation of the pseudo-population which can require a lot of space when the population is very large. We have also highlighted the difficulties involved in variance estimation and the construction of certain confidence intervals, such as the popular percentile bootstrap intervals, when the pseudo-population method involves a random pseudo-population as opposed to a fixed one. Direct bootstrap methods involve independent sampling (in a context where sampling without replacement is usually part of the sampling plan) and tuning parameters must be set to reflect the sampling variability of the sampling plan. While all choices of the tuning parameters will lead to the same variance estimate for the estimator of population total, bootstrap confidence intervals will give different results, even for the population total. Similarly, many of the bootstrap weights methods are constructed to provide the same estimator of variance for the total (as well as the same bootstrap mean of the estimator), but the bootstrap distribution, and therefore confidence intervals, will depend on the actual method.

In fact, with few exceptions, all bootstrap methods, even the class of pseudo-population ones, can be described as bootstrap weights methods in that the original survey weights can be modified to take into account how often one of the original observation appears in the bootstrap sample. Hence, these methods can also be used in public files provided by statistical agencies. On the other hand, it is not clear for certain of these methods that an efficient algorithm exists to provide the bootstrap weights without actually creating the pseudo-population and taking the sample from it to count the number of appearances of each original observation.

We have also seen that software availability of adequate sample survey bootstrap methods is still lagging. The use of existing columns of bootstrap weights created by statistical agencies for researchers is relatively simple. But even then, as we have seen, it may be important to know more about the method that created the bootstrap weights as otherwise the computation of a bootstrap estimate of variance or the construction of a confidence interval may be incorrect if a pseudo-population method with a random pseudo-population is used. As for the construction of bootstrap weights, relatively little is available, appropriate documentation is lacking, finite sample correction is often not available, and there is basically nothing readily available for unequal probability sampling design.

Sample surveys often suffer from missing observations, either through unit non-response or through item non-response. In the first case, one possibility to address the problem is to reweight the units who have responded. But for unit non-response, the solution often involves imputation. It is important to understand that the bootstrap weights that researchers can use in public data file only reflect the sampling variability of the selected units, not the extra variability due to the imputation. This extra variability, which depends on the non-response mechanism and the imputation method, requires special care.

While a lot has been accomplished in bootstrap research for survey sampling, much work remains. For instance, many of the simulations in the case of quantiles such as the median have shown relatively poor results whether it be for variance estimation or for confidence intervals; see e.g., Sitter (1992b) for stratified simple random sampling or Saigo (2010) for three-stage stratified simple random sampling. One possible explanation may be the emphasis on requesting that the bootstrap method match the first two sample moments for the estimation of the total which behaves very differently from quantiles. More work is needed for such estimators.

References

Aitkin, M. (2008). Applications of the bayesian bootstrap in finite population inference. *Journal of Official Statistics 24*(1), 21–51.

Antal, E. and Y. Tillé (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association 106*(494), 534–543. MR2847968

Antal, E. and Y. Tillé (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics 29*(5), 1345–1363. MR3266062

Barbe, P. and P. Bertail (1995). *The weighted bootstrap, Lecture notes in statistics*, Volume 98. Springer-Verlag, New York. MR2195545

Barbiero, A., G. Manzi, and F. Mecatti (2015). Bootstrapping probability-proportional-to-size samples via calibrated empirical population. *Journal of Statistical Computation and Simulation 85*(3), 608–620. MR3275468

Barbiero, A. and F. Mecatti (2010). Bootstrap algorithms for variance estimation in πps sampling. In *Complex data modeling and computationally intensive statistical methods*, pp. 57–69. Springer.

Beaumont, J.-F. and A.-S. Charest (2012). Bootstrap variance estimation with survey data when estimating model parameters. *Computational Statistics and Data Analysis 56*(12), 4450–4461. MR2957885

Beaumont, J.-F. and Z. Patak (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review 80*(1), 127–148. MR2990349

Berger, Y. G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika 94*(4), 953–964. MR2416801

Berger, Y. G. and C. J. Skinner (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 79–89. MR2136640

Bertail, P. and P. Combris (1997). Bootstrap généralisé d'un sondage. *Annales d'économie et de statistique 46*, 49–83. MR1478018

Bickel, P. J. and D. A. Freedman (1983). Asymptotic normality and the bootstrap in stratified sampling. *Unpublished manuscript. Department of Statistics, University of California, Berkeley.* MR0740906

Bickel, P. J. and D. A. Freedman (1984). Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics 12*(2), 470–482. MR0740906

Binder, D. A. (2011). Estimating model parameters from a complex survey under a model-design randomization framework. *Pakistan Journal of Statistics 27*(4), 371–390. MR2919725

Booth, J. G., R. W. Butler, and P. Hall (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association 89*(428), 1282–1289. MR1310222

Campbell, C. and A. D. Little (1980). A different view of finite population estimation. In *Proceedings of the Section on Survey Research Methods*, pp. 319–324.

Canty, A. J. and A. C. Davison (1999). Resampling-based variance estimation for labour force surveys. *Journal of the Royal Statistical Society, Series D: The Statistician 48*, 379–391.

Carota, C. (2009). Beyond objective priors for the bayesian bootstrap analysis of survey data. *Journal of Official Statistics 25*(3), 405–413.

Chao, M. T. and S.-H. Lo (1985). A bootstrap method for finite population. *Sankhyā: The Indian Journal of Statistics, Series A 47*, 399–405. MR0863733

Chao, M. T. and S.-H. Lo (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica 4*(2), 389–406. MR1309420

Chaudhuri, A. and A. Saha (2004). Extending sitter's mirror-match bootstrap to cover rao-hartley-cochran sampling in two-stages with simulated illustrations. *Sankhya: The Indian Journal of Statistics 66*(4), 791–802. MR2205821

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

Chipperfield, J. and J. Preston (2007). Efficient bootstrap for business surveys. *Survey Methodology 33*(2), 167–172.

Davison, A. and D. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR1478673

Deville, J. C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology 25*(2), 193–203.

Durbin, J. (1959). A note on the application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika 46*(3-4), 477–480. MR0110146

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics 7*(1), 1–26. MR0515681

Escobar, E. L. and Y. G. Berger (2013). A jackknife variance estimator for self-weighted two-stage samples. *Statistica Sinica 23*(2), 595–613. MR3086648

Fay, R. E. (1991). A design-based perspective on missing data variance. In *Proceedings of the 1991 Annual Research Conference, US Bureau of the census*, pp. 429–440.

Fuller, W. A. (2009). *Sampling statistics*. Wiley, New York.

Funaoka, F., H. Saigo, R. R. Sitter, and T. Toida (2006). Bernoulli bootstrap for stratified multistage sampling. *Survey Methodology 32*(2), 151–156.

Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 181–184.

Gupta, V. K. and A. K. Nigam (1987). Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika 74*(4), 735–742. MR0919841

Gurney, M. and R. S. Jewett (1975). Constructing orthogonal replications for variance estimation. *Journal of the American Statistical Association 70*(352), 819–821.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In C. Rao and D. Pfeffermann (Eds.), *Handbook of Statistics 29A, Sample Sur-*

veys: Design, Methods and Applications, pp. 215–246. Elsevier. MR2654640

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.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685. MR0053460

Isaki, C. T. and W. A. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association 77*(377), 89–96. MR0648029

Jones, H. L. (1974). Jackknife estimation of functions of stratum means. *Biometrika 61*(2), 343–348. MR0394966

Kim, J. K., A. Navarro, and W. A. Fuller (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association 101*(473), 312–320. MR2268048

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal, The: The official journal on Stata and statistics 10*(2), 165–199.

Kott, P. S. (1998). Using the delete-a-group jackknife variance estimator in practice. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 763–768.

Kott, P. S. (2001). The delete-a-group jackknife. *Journal of Official Statistics 17*(4), 521–526.

Kovacevic, M. S., R. Huang, and Y. You (2006). Bootstrapping for variance estimation in multi-level models fitted to survey data. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 3260–3269.

Kovar, J. G., J. N. K. Rao, and C. F. J. Wu (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics 16, Supplement*, 25–45. MR0997120

Krewski, D. and J. N. K. Rao (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics 9*(5), 1010–1019. MR0628756

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science 18*(2), 199–210. MR2019788

Lo, A. Y. (1991). Bayesian bootstrap clones and a biometry function. *Sankhyā: The Indian Journal of Statistics, Series A 53*(3), 320–333. MR1189775

Lumley, T. (2010). *Complex surveys: a guide to analysis using R*. Wiley, Hoboken, NJ.

Lumley, T. (2014). survey: analysis of complex survey samples. R package version 3.30.

Mashreghi, Z., C. Léger, and D. Haziza (2014). Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *The Canadian Journal of Statistics 42*(1), 142–167. MR3181587

Mason, D. M. and M. A. Newton (1992). A rank statistics approach to the consistency of a general bootstrap. *The Annals of Statistics 20*(3), 1611–1624. MR1186268

McCarthy, P. J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute 37*(3), 239–264.

McCarthy, P. J. and C. B. Snowden (1985). The bootstrap and finite population sampling. *Vital and Health Statistics*, Series 2, No. 95. DHHS Publication No. (PHS) 85–1369. Public Health Service. Washington. U.S. Government Printing Office.

Miller, R. G. (1974). The jackknife-a review. *Biometrika 61*(1), 1–15. MR0391366

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology 35*(2), 227–234.

Preston, J. and J. Chipperfield (2002). Using a generalised estimation methodology for ABS business surveys. *Methodology Advisory Committee, ABS, Belconnen, Australia (available at www.abs.gov.au).*

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika 43*(3-4), 353–360. MR0081040

Ranalli, M. G. and F. Mecatti (2012). Comparing recent approaches for bootstrapping sample survey data: A first step toward a unified approach. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 4088–4099.

Rao, J. N. K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association 3*, 173–180. MR0201035

Rao, J. N. K., H. O. Hartley, and W. G. Cochran (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society. Series B (Methodological) 24*(2), 482–491. MR0148196

Rao, J. N. K. and J. Shao (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association 91*(433), 343–348. MR1394090

Rao, J. N. K. and J. Shao (1999). Modified balanced repeated replication for complex survey data. *Biometrika 86*(2), 403–415. MR1705398

Rao, J. N. K. and C. F. J. Wu (1985). Inference from stratified samples: second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association 80*(391), 620–630. MR0803259

Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association 83*(401), 231–241. MR0941020

Rao, J. N. K., C. F. J. Wu, and K. Yue (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology 18*(2), 209–217.

Rosén, B. (1997). Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference 62*(2), 135–158. MR1468158

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592. MR0455196

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics 1*(4), 381–397.

Rust, K. F. and J. N. K. Rao (1996). Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research 5*(3), 283–310.

Saigo, H. (2010). Comparing four bootstrap methods for stratified three-stage sampling. *Journal of Official Statistics 26*(1), 193–207.

Saigo, H., J. Shao, and R. R. Sitter (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology 27*(2), 189–196.

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika 54*(3-4), 499–513. MR0223051

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology 33*(2), 99–119.

Särndal, C.-E., B. Swensson, and J. H. Wretman (1997). *Model assisted survey sampling*. Berlin; New York: Springer-Verlag Inc.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science 18*(2), 191–198. MR2019787

Shao, J. and Y. Chen (1998). Bootstrapping sample quantiles based on complex survey data under hot deck imputation. *Statistica Sinica 8*(4), 1071–1085. MR1666229

Shao, J. and J. N. K. Rao (1993). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā: The Indian Journal of Statistics, Series B 55*(3), 393–414. MR1319142

Shao, J. and R. R. Sitter (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association 91*(435), 1278–1288. MR1424624

Shao, J. and P. Steel (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association 94*(445), 254–265. MR1689230

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics, New York. MR1351010

Shao, J. and C. F. J. Wu (1989). A general theory for jackknife variance estimation. *The Annals of Statistics 17*(3), 1176–1197. MR1015145

Shao, J. and C. F. J. Wu (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics 20*(3), 1571–1593. MR1186266

Sitter, R. R. (1992a). Comparing three bootstrap methods for survey data. *The Canadian Journal of Statistics 20*(2), 135–154. MR1183077

Sitter, R. R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association 87*(419), 755–765. MR1185197

Sitter, R. R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika 80*(1), 211–221. MR1225226

Tukey, J. W. (1958). Bias and confidence in not quite large samples. Abstract. *The Annals of Mathematical Statistics 29*, 614.

Wang, Z. and M. E. Thompson (2012). A resampling approach to estimate variance components of multilevel models. *The Canadian Journal of Statistics 40*(1), 150–171. MR2896935

Wolter, K. M. (2007). *Introduction to Variance Estimation*. Springer Series in Statistics, New York.

Wu, C. F. J. (1991). Balanced repeated replications based on mixed orthogonal arrays. *Biometrika 78*(1), 181–188. MR1118243