Comparing Recent Approaches For Bootstrapping Sample Survey Data: A First Step Towards A Unified Approach

M. Giovanna Ranalli* Fulvia Mecatti†

Abstract

Bootstrap algorithms are simple and appealing solutions for variance estimation under a complex sampling design, however, they must account for the non-iid nature of data. Literature about bootstrapping finite population samples appears to have developed according to two major approaches. A more practical *ad-hoc* approach refers to the so-called scaling problem and is based on a data-rescaling so that, in the linear case, the resulting bootstrap estimate for the variance perfectly matches the analytic variance estimate. A more fundamental *plug-in* approach is based on the mimicking bootstrap principle and on the bootstrap population created on the basis of (original) sample data. Recent proposals suggest a direct bootstrap matching the linear case variance but avoiding any data scaling under mixed re-sampling designs. In this paper, a new perspective to the bootstrap population plug-in approach is provided that avoids the physical reconstruction of the bootstrap population. Basic sampling designs, both with and without replacement as well as unequal probability designs are considered. Focusing on probability-proportional-to-size sampling, a simulation study is conducted that compares all the approaches considered.

Key Words: Bootstrap principles, Conditional Poisson design, non-central Hypergeometric distribution, Probability-proportional-to-size design, Pseudo-population, Variance estimation.

1. Introduction

Bootstrap algorithms are simple and general tools for (a) assessing estimators' accuracy via variance estimation, and (b) producing confidence intervals and p-values. Bootstrap applies to finite samples and provides numerical solutions for non-standard situations so that it is particularly appealing when dealing with finite populations and complex sampling designs.

We focus on a general sampling design where each population unit is assigned a specific probability to be included in the sample, not necessarily equal, and the sample contains distinct units only. It is of particular practical interest in this framework the without replacement probability proportional to size – πps – sampling, where the inclusion probabilities are set proportional to an available positive auxiliary variable. πps sampling is extensively used in large scale survey for it has increasing efficiency potential as the relation between the study and the auxiliary variable approaches proportionality versus a conventional (equal probability) simple random sampling - SRS.

However, crucial survey objectives as in (a) and (b) above are challenging under a πps sampling even in the simplest linear case, i.e. estimation of means and totals, essentially for computational reasons and become unmanageable for more complex non-linear cases, e.g. estimation of quantiles, multi-dimensional indicators etc. A bootstrap solution appears then appropriate. However, since the original Efron's bootstrap (Efron, 1979) applies to independent and identical distributed – iid – sample data suitable adaptations are needed in order to account for the non-iid nature of data due to the complexity of the sampling design.

*Department of Economics, Finance and Statistics, University of Perugia, Via Pascoli, 06123, Perugia, Italy, giovanna.ranalli@stat.unipg.it

†Department of Statistics, University of Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy, fulvia.mecatti@unimib.it

Literature about bootstrapping finite population (complex) samples appears to have developed according to two major approaches, which we will refer to as *ad-hoc* and *plug-in*, respectively (see e.g. Presnell and Booth, 1994; Chauvet, 2007).

1. The ad-hoc approach is based on iid re-sampling, as for the original – naïve – bootstrap, and requires the re-scaling of sample data in such a way that in the linear case the final bootstrap estimate for the variance perfectly matches the analytic variance estimate. This more practical approach includes for instance the with-replacement bootstrap (McCarthy and Snowden, 1985) the rescaling bootstrap (Rao and Wu, 1988), the mirror-match bootstrap (Sitter, 1992), the generalized bootstrap based on weighting (Bertail and Combris, 1997; Beaumont and Patak, 2012). The recently proposed direct bootstrap (Antal and Tillé, 2011), though based on variance matching, involves non-iid re-sampling and it does not require data re-scaling.

2. The plug-in approach is more fundamental and is based on bootstrap principles such as the mimicking principle (Hall, 1992) and the bootstrap population (see Gross, 1980; Chao and Lo, 1985; Booth *et al.*, 1994 for equal probability sampling designs; Holmberg, 1998; Chauvet, 2007; Barbiero and Mecatti, 2009 for unequal probability sampling designs).

We focus here on approach 2. Besides being consistent with bootstrap principles and foundations, it appears preferable for handling πps sampling. In fact, the variance matching ad-hoc approach might be problematic to apply owing to the many alternative variance estimators available for the linear case, either exact or approximate, thus requiring problem-dependent arbitrary choices to be justified. A bootstrap population is a pseudo or empirical population built up by using sample data only, and assumed to estimate the unknown parent population. According to the mimicking principle, bootstrap samples (i.e. the re-sampling result) are selected from this estimated population with the same sample size as the original sample and by mimicking the original sampling design to the largest extent.

Algorithms based on the bootstrap population are proved to ensure second order accuracy as for the Efron's iid bootstrap (Booth *et al.*, 1994). However, the reconstruction of the bootstrap population may be cumbersome and repeated πps selections from it extremely time consuming. This work aims at developing a method to drive a completely plug-in bootstrap inference by directly re-sampling from the sample, with no need to physically reconstruct the bootstrap population. We will consider a number of popular sampling designs, both with and without replacement, with equal and unequal probabilities. Empirical evidence from a limited simulation study is also provided, in the case of πps sampling and for estimating different linear, semi-linear and non-linear finite population parameters of interest. The simulation has the main objective of evaluating the performance and applicability of the proposed methodology and to compare it with recent competitors.

2. Notation

Let $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$ be a finite population of size N and let a random sample be represented by a random vector

$$\boldsymbol{S} = (S_1, \ldots, S_k, \ldots, S_N),$$

where S_k indicates the number of times unit k is selected in the sample. For sampling without replacement – WOR – S_k is the sample membership indicator taking value 1 if unit k is included in the sample and 0 otherwise (see Traat et al., 2004; Tillé, 2006, for examples of use of this notation). For designs with fixed sample size $\sum_{k=1}^{N} S_k = n$. Let $\boldsymbol{s}$

be a realization of $\boldsymbol{S}$. Then, the sampling design is the (multivariate) probability distribution $p(\boldsymbol{s}) = P(\boldsymbol{S} = \boldsymbol{s})$, with $\sum_{\boldsymbol{s}} p(\boldsymbol{s}) = 1$. Inclusion probabilities are given by the expectation of the product of the components of any sub-vector of $\boldsymbol{S}$, for instance $\pi_k = E(S_k)$ and $\pi_{k\ell} = E(S_k S_\ell)$ for first and second order inclusion probabilities, respectively.

Let $Y = \sum_{k=1}^{N} y_k$ be the total of a study variable y with design-unbiased estimator

$$\hat{Y} = \sum_{k=1}^{N} S_k y_k/\pi_k = \sum_{k\in\mathcal{S}} y_k/\pi_k = \sum_{k\in\mathcal{S}} d_k y_k,$$

where $\mathcal{S} = \{k \text{ s.t. } S_k > 0\}$ is the set of labels selected in the sample (possibly repeated). For any WOR sampling design, its variance

$$V(\hat{Y}) = \sum_{k=1}^{N}\sum_{\ell=1}^{N} \frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}(\pi_{k\ell} - \pi_k\pi_\ell)$$

can be estimated by either the familiar Horvitz-Thompson or Sen-Yates-Grundy variance estimator, both design-unbiased and available with well-known analytic and closed formulae.

According to this notation, a general finite population bootstrap algorithm for variance estimation can be illustrated as follows. A bootstrap sample $\boldsymbol{s}^*$ – sometimes named re-sample – informed by (original) sample data only in assorted frequency, is produced by generating a realization of the random vector $\boldsymbol{S}^* = (S_k^*,\ k \in \mathcal{S})$ (re-sampling step). A replicate of the estimator is then computed over the bootstrap sample (replication step), for instance in the linear case $\hat{Y}^* = \sum_{k\in\mathcal{S}} S_k^* y_k/\pi_k$. The process is iterated a large number of times, say C, to obtain the bootstrap distribution, for instance in the linear case $\hat{Y}_c^*$, for $c = 1, \ldots, C$. The bootstrap distribution is assumed as a $C-$run Monte Carlo estimate of the actual (usually unknown) estimator distribution and then used to produce the bootstrap estimate of interest, the bootstrap variance estimate being for instance in the linear case

$$V^*(\hat{Y}^*) = \frac{1}{C-1}\sum_{c=1}^{C}(\hat{Y}_c^* - \bar{\hat{Y}^*})^2.$$

Bootstrap algorithms under the ad-hoc approach as described at point 1. of the Introduction, provide a perfect variance matching in the linear case, namely $V^*(\hat{Y}^*) = \hat{V}(\hat{Y})$, in addition to the traditional matching $E^*(\hat{Y}^*) = \hat{Y}$, which we will refer to as bootstrap unbiasedness as featuring the iid Efron's bootstrap already. This is accomplished either by an iid re-sampling of possibly different size $n^* \neq n$ joined with a data-rescaling as in the Rao-Wu bootstrap, or by performing the re-sampling under a particular mixture of designs as for the recently proposed direct bootstrap (Antal and Tillé, 2011).

On the other side, algorithms based on the bootstrap population as mentioned at point 2. of the Introduction, enforce the mimicking principle to a larger extent. A bootstrap-population step is included on top of the algorithm where a set $\mathcal{U}^*$ is created by replicating d_k^* times each unit $k \in \mathcal{S}$. $\mathcal{U}^*$ is assumed as an estimate of the unknown actual population $\mathcal{U}$ - from which the original sample $\mathcal{S}$ is selected - and thus used to perform the re-sampling. According to the mimicking principle, $\mathcal{U}^*$ should copy known features of $\mathcal{U}$ and the re-sampling design $p^*(\boldsymbol{s}^*|\mathcal{U}^*)$ should be mirroring the original design $p(\boldsymbol{s})$. For instance the WOR bootstrap (Gross, 1980; Chao and Lo, 1985; Booth *et al.*, 1994) applying to SRS, involves the choice $d_k^* = N/n$ ensuring a bootstrap population with the same size $N^* = \sum_{k=1}^{n} d_k^* = N$ and a WOR re-sampling from $\mathcal{U}^*$ of size $n^* = n$. For a πps design the choice $d_k^* = 1/\pi_k$ ensures the matching of the total of the auxiliary size variable,

say x, namely $X^* = \sum_{k\in\mathcal{S}} d_k^* x_k = X = \sum_{k=1}^N x_k$, and a πps re-sampling is suggested (Holmberg, 1998; Chauvet, 2007; Barbiero and Mecatti, 2009).

This approach obeys to a real plug-in principle as for the Efron's iid Bootstrap (Booth et al., 1994; Chauvet, 2007) and it is not just motivated by a forced adjustment on the first two moments of the estimator of the linear case. However, the plug-in principle has the drawback of being resource-consuming and potentially cumbersome in the reconstruction of the pseudo-population as well as in re-sampling from it by mirroring the original sampling design, thus substantially limiting its application. In the next section we propose a methodology for bootstrapping under the bootstrap population approach by re-sampling directly from the original sample $\mathcal{S}$, in fact skipping both the physical reconstruction of $\mathcal{U}^*$ and the re-sampling from it.

3. Methodology proposed for implementing the Plug-in approach

A number of sampling designs of increasing complexity will be considered. For simplicity we will assume that the bootstrap weights d_k^* are integer for all $k \in \mathcal{S}$. This is a strong assumption, rarely realized in real applications. At the end of this section a discussion about how to deal with this issue is provided.

3.1 Simple random sampling with replacement

When the original sample from $\mathcal{U}$ is obtained by SRS with replacement, $\boldsymbol{S}$ has a Multinomial distribution, i.e. $\boldsymbol{S} \sim \mathrm{M}(n; \frac{1}{N}, \ldots, \frac{1}{N})$ defined as

$$p(\boldsymbol{s}) = n! \prod_{k=1}^{N} \frac{(1/N)^{S_k}}{S_k!}$$

with $S_k \in \{0, 1, \ldots, n\}$. The bootstrap population $\mathcal{U}^*$ involves the familiar SRS choice $d_k^* = N/n$, i.e. each sampled unit $k \in \mathcal{S}$ is replicated N/n times leading to $N^* = N$ under the integer assumption. Re-sampling from $\mathcal{U}^*$ mimicking the original sampling design gives $\boldsymbol{S}^* \sim \mathrm{M}(n; \frac{1}{N^*}, ..., \frac{1}{N^*})$ and

$$p^*(\boldsymbol{s}^*|\mathcal{U}^*) = n! \prod_{k\in\mathcal{U}^*} \frac{(1/N^*)^{S_k^*}}{S_k^*!}$$

with $S_k^* \in \{0, 1, \ldots, n\}$, $k \in \mathcal{U}^*$.

Note that $\mathcal{U}^*$ is made by means of only n distinct units each one with frequency N/n; as a consequence re-sampling from $\mathcal{U}^*$ with probabilities $1/N^*$, $\forall k \in \mathcal{U}^*$ is equivalent to re-sampling from $\mathcal{S}$ with probabilities $(1/N^*)(N^*/n) = 1/n$, i.e. $\boldsymbol{S}^*|\mathcal{S} \sim \mathrm{M}(n; \frac{1}{n}, \ldots, \frac{1}{n})$ and

$$p^*(\boldsymbol{s}^*|\mathcal{S}) = n! \prod_{k\in\mathcal{S}} \frac{(1/n)^{S_k^*}}{S_k^*!}$$

with $S_k^* \in \{0, 1, \ldots, n\}$, $k \in \mathcal{S}$. Bootstrap unbiasedness is automatically provided in this case since $E^*(S_k^*|\mathcal{S}) = 1$. This is indeed a general property of the methodology that holds in all subsequent cases. In addition, in this case, it can be shown that the bootstrap variance estimate results $V^*(\hat{Y}^*) = \frac{n-1}{n}\hat{V}(\hat{Y})$. This is, indeed, the original iid Efron's Bootstrap.

3.2 Sampling with unequal probabilities and with replacement

When selecting with replacement and unequal probabilities, each population unit k has attached a fixed selection probability p_k and each element is replaced in the population after it

is drawn. The resulting multivariate distribution of $\boldsymbol{S}$ is Multinomial, i.e. $\boldsymbol{S} \sim \mathrm{M}(n; p_1, \ldots, p_N)$, so that

$$p(\boldsymbol{s}) = n! \prod_{k=1}^{N} \frac{p_k^{S_k}}{S_k!} \tag{1}$$

with $S_k \in \{0, 1, \ldots, n\}$. Again, assuming $d_k^* = 1/np_k$ is integer, $\mathcal{U}^*$ is made of $N^* = \sum_{k \in \mathcal{S}} d_k^*$ units, in which each unit $k \in \mathcal{S}$ is replicated d_k^* times. Then re-sampling from $\mathcal{U}^*$ is such that $\boldsymbol{S}^* \sim \mathrm{M}(n; p_k, k \in \mathcal{U}^*)$ and

$$p^*(\boldsymbol{s}^*|\mathcal{U}^*) = n! \prod_{k \in \mathcal{U}^*} \frac{(p_k)^{S_k^*}}{S_k^{*!}}$$

with $S_k^* \in \{0, 1, \ldots, n\}$, $k \in \mathcal{U}^*$.

Still note that the bootstrap population $\mathcal{U}^*$ is effectively made of n distinct units each with frequency d_k^*. Consequentely $p^*(\boldsymbol{s}^*|\mathcal{U}^*)$ is equivalent to re-sample with replacement from $\mathcal{S}$ with probabilities $d_k^* p_k = 1/n$. That is, $\boldsymbol{S}^*|\mathcal{S} \sim \mathrm{M}(n; \frac{1}{n}, \ldots, \frac{1}{n})$ and

$$p^*(\boldsymbol{s}^*|\mathcal{S}) = n! \prod_{k \in \mathcal{S}} \frac{(1/n)^{S_k^*}}{S_k^{*!}}$$

with $S_k^* \in \{0, 1, \ldots, n\}$, $k \in \mathcal{S}$. It can be shown that $V^*(\hat{Y}^*) = \frac{n-1}{n} \hat{V}(\hat{Y})$. This shows that when sampling with replacement, even if the original design is with unequal probabilities, the pseudo-population approach reduces to re-sampling directly from the sample with equal probabilities, i.e. using the naïve iid bootstrap. Antal and Tillé (2011) reach a similar conclusion when applying their direct bootstrap methodology.

3.3 Poisson sampling

Under Poisson sampling, the sample selection is random-size and list-sequential by performing as many independent trials as the population size, each with probability π_k. Therefore, for every population unit $k \in \mathcal{U}$ the sample membership indicator has Bernoulli distribution, i.e. $S_k \sim \mathrm{Be}(\pi_k)$, so that

$$p(\boldsymbol{s}) = \prod_{k=1}^{N} \pi_k^{S_k} (1 - \pi_k)^{1 - S_k} \tag{2}$$

with $S_k \in \{0, 1\}$. Using the plug-in principle, the bootstrap population $\mathcal{U}^*$ is built up by replicating $d_k^* = d_k = \pi_k^{-1}$ times each sampled unit $k \in \mathcal{S}$ and the re-sampling from it is such that $S_k^* \sim \mathrm{Be}(\pi_k)$ for $k \in \mathcal{U}^*$, that is

$$p^*(\boldsymbol{s}^*|\mathcal{U}^*) = \prod_{k \in \mathcal{U}^*} \pi_k^{S_k^*} (1 - \pi_k)^{1 - S_k^*}$$

with $S_k^* \in \{0, 1\}$, $k \in \mathcal{U}^*$.

Nevertheless, this procedure is equivalent to re-sampling from $\mathcal{S}$ by generating S_k^* independently for every $k \in \mathcal{S}$ distinct in $\mathcal{U}^*$ with frequency d_k, i.e. with Binomial distribution $S_k^*|\mathcal{S} \sim \mathrm{Bin}(d_k, \pi_k)$. This gives

$$p^*(\boldsymbol{s}^*|\mathcal{S}) = \prod_{k \in \mathcal{S}} \binom{d_k}{\pi_k} \pi_k^{S_k^*} (1 - \pi_k)^{d_k - S_k^*}$$

with $S_k^* \in \{0, 1, ..., d_k\}$, $k \in \mathcal{S}$. This equivalence has also been noted by Beaumont and Patak (2012), who prove that not only $V^*(\hat{Y}^*) = \hat{V}(\hat{Y})$, but that this procedure matches also the third design moment of the sampling error.

3.4 Simple random sampling without replacement

When units are selected with equal probabilities and WOR, the random vector $\boldsymbol{S}$ representing the (original) sample selection has multivariate Hypergeometric distribution. By referring to the classical urn representation, we have n selections out of N balls of distinct colors (population units) each with maximal possible selection count equal to 1 (WOR selection). Thus $\boldsymbol{S} \sim$ Multi.Hyperg$(n; 1, ..., 1)$ and

$$p(\boldsymbol{s}) = \binom{N}{n}^{-1} \prod_{k=1}^{N} \binom{1}{S_k} = \binom{N}{n}^{-1}$$

with $S_k \in \{0, 1\}$, $k \in \mathcal{U}$ (see also Traat et al., 2004 for details on this). As for the SRS with replacement case in Section 3.1, the bootstrap population $\mathcal{U}^*$ is built by replicating $d_k^* = N/n$ times each sampled unit $k \in \mathcal{S}$ and the re-sampling vector mimics the original WOR selection, i.e. $\boldsymbol{S}^* \sim$ Multi.Hyperg$(n; 1, ..., 1)$. Hence

$$p^*(\boldsymbol{s}^*|\mathcal{U}^*) = \binom{N^*}{n}^{-1}$$

with $S_k^* \in \{0, 1\}$, $k \in \mathcal{U}^*$.

Note that $\mathcal{U}^*$ is indeed an urn comprising n balls of distinct colors each with frequency N/n so that re-sampling from $\mathcal{U}^*$ is equivalent to re-sample from $\mathcal{S}$ under the *working* re-sampling vector $\boldsymbol{S}^*|\mathcal{S} \sim$ Multi.Hyperg$(n; \frac{N}{n}, ..., \frac{N}{n})$ which gives

$$p^*(\boldsymbol{s}^*|\mathcal{S}) = \binom{N}{n}^{-1} \prod_{k \in \mathcal{S}} \binom{N/n}{S_k}$$

with $S_k^* \in \{0, 1, \ldots, N/n\}$, $k \in \mathcal{S}$. It can be shown that this leads to the familiar WOR Bootstrap variance estimate for the linear case (Chao and Lo, 1985)

$$V^*(\hat{Y}^*) = \frac{N}{N-1} \frac{n-1}{n} \hat{V}(\hat{Y}).$$

3.5 πps sampling

We finally consider a fixed-size πps design with inclusion probability exactly proportional to a known (positive) auxiliary variable x, i.e. $\pi_k = nx_k/X$, with $X = \sum_{k=1}^{N} x_k$. This figures a more complex case than in previous sections, in fact including a large collection of different designs each providing a particular set of joint inclusion probabilities (see e.g. Brewer and Hanif, 1983; Tillé, 2006, Ch. 6, 7). Each of this fixed-size (exactly) πps design therefore induces a different joint multivariate distribution for the sampling vector $\boldsymbol{S}$. We will consider here the special though relevant case of Conditional Poisson sampling as starting point to illustrate our πps-Bootstrap methodology and to discuss other possibilities.

The Conditional Poisson design is essentially a Poisson design in which the sampling size is fixed to be equal to n. This can be achieved for instance by rejection (see Tillé, 2006, for a set of different algorithms to select Conditional Poisson samples). By conditioning on a fixed sample size, the basic practical disadvantage of Poisson sampling is removed while maintaining the appealing simplicity. The joint distribution of $\boldsymbol{S}$ can be obtained by suitably conditioning the probability distribution of the Poisson design in (2), i.e.

$$p(\boldsymbol{s}) = C_1 \prod_{k=1}^{N} \pi_k^{S_k} (1 - \pi_k)^{1 - S_k} \quad \text{if } \sum_{k=1}^{N} S_k = n, \tag{3}$$

where $S_k \in \{0,1\}$, $k \in \mathcal{U}$ and C_1 is the normalizing constant (see also Traat et al., 2004). Note that under a Conditional Poisson design the probability distribution $p(\boldsymbol{s})$ can equivalently be derived as a conditional Multinomial distribution, given that $S_k \leq 1$ and $\sum_{k=1}^{N} S_k = n$ (for connections between Poisson, Multinomial and Conditional Poisson designs see Tillé, 2006, Chap. 5). In particular, if $\tilde{\boldsymbol{S}} \sim M(n; p_1, p_2, \ldots, p_N)$ and $\tilde{S}_k \leq 1$, then from (1)

$$p(\tilde{\boldsymbol{s}}|\tilde{S}_k \leq 1, \sum_{k=1}^{N} \tilde{S}_k = n) = p(\boldsymbol{s}) = C_2 \prod_{k=1}^{N} p_k^{S_k}, \quad \text{if } \sum_{k=1}^{N} S_k = n. \tag{4}$$

Note that if $p_k \propto \pi_k/(1-\pi_k)$, then (3) and (4) coincide.

With the natural choice $d_k^* = d_k = \pi_k^{-1}$ (Holmberg, 1998) for constructing the bootstrap population $\mathcal{U}^*$ and under a (mimicking) Conditional Poisson re-sampling we have:

$$p^*(\boldsymbol{s}^*|\mathcal{U}^*) = C_3 \prod_{k \in \mathcal{U}^*} \pi_k^{*S_k^*}(1-\pi_k^*)^{1-S_k^*}, \quad \text{if } \sum_{k \in \mathcal{U}^*} S_k^* = n, \tag{5}$$

where $\pi_k^* = n x_k / X^*$ and $X^* = \sum_{k \in \mathcal{U}^*} x_k = \sum_{k \in \mathcal{S}} d_k^* x_k$ is the bootstrap auxiliary total.

As in Section 3.4, $\mathcal{U}^*$ can be thought as an urn comprising N^* balls of n distinct colors each with frequency d_k^*. Therefore the re-sampling can be associated with the experiment of taking colored balls from $\mathcal{U}^*$ at random and without replacement. However, differently from the SRS case, each unit has now a specific and possibly different probability ($\propto x_k$) of being selected, leading to balls of one color that have a higher probability of being taken than balls of another color. This is named a *biased urn* setting and the number of balls drawn of each color follows a non-central multivariate Hypergeometric distribution (Johnson *et al.*, 1997, Chap. 39). The distribution, because of the different probability each ball is given, depends on how the balls are taken from the urn. In the literature, two different probability distributions are known as non central multivariate Hypergeometric: Wallenius' and Fisher's. The former is obtained if n balls are taken one by one, while the latter if balls are taken still WOR and independently of each other (see Fog, 2008, for a detailed distinction between the two). As a consequence, re-sampling from $\mathcal{U}^*$ by generating from $p^*(\boldsymbol{s}^*|\mathcal{U}^*)$ in (5) is equivalent to re-sampling directly from $\mathcal{S}$ by generating from a Fisher non-central Hypergeometric distribution, i.e.

$$\boldsymbol{S}^*|\mathcal{S} \sim \text{F.nc.Multi.Hyperg}(n; d_k; \omega_k^*; \text{for } k \in \mathcal{S}),$$

where n is the number of colors (distinct units $k \in \mathcal{S}$ appearing in $\mathcal{U}^*$), d_k is the frequency of color k in $\mathcal{U}^*$ and ω_k^* is the *weight* associated to balls of color k, so that the probability that a particular ball is sampled at a given draw is proportional to its weight. Then

$$p^*(\boldsymbol{s}^*|\mathcal{S}) = C_4 \prod_{k \in \mathcal{S}} \binom{d_k}{S_k^*} \omega_k^{*S_k^*} \tag{6}$$

where C_4 is a normalizing constant given by the sum over those draws for which $\sum_{k \in s} S_k^* = n$. Note that (6) can be obtained from (5) by setting weight ω_k^* proportional to the odds of unit k, i.e. if $\omega_k^* \propto \pi_k^*/(1-\pi_k^*)$.

Draw-by-draw WOR (fixed-size exactly) πps designs can be handled similarly but by referring to Wallenius' non-central Hypergeometric distribution. Further research concerning this most complex case is needed.

3.6 Non-integer d_k's

As mentioned at the beginning of this section, so far we have assumed integer weights $d_k^* = d_k$, $k \in \mathcal{S}$ for constructing the bootstrap population $\mathcal{U}^*$. We have made this assumption also for illustrating the equivalence between re-sampling from it by mimicking the original sampling design and re-sampling directly from the (original) sample $\mathcal{S}$ by generating from a suitable working probability distribution $p^*(\boldsymbol{s}^*|\mathcal{S})$ clearly depending on d_k^*. Such integer assumption is rarely fulfilled in real applications even in the simpler constant case $d_k^* = N/n$ and become unrealistic in the general case $d_k^* = \pi_k^{-1}$, for which it should hold for all $k \in \mathcal{S}$. According to a recurring suggestion for dealing with non-integer d_k^*, a further randomization step is often added on top of the bootstrap algorithm – as described in Section 1 – producing a set of integer weights d_k^* by means of n independent Bernoulli trials. In particular,

$$d_k^* = \begin{cases} \lfloor d_k \rfloor & \text{with probability } 1 - (d_k - \lfloor d_k \rfloor) \\ \lfloor d_k \rfloor + 1 & \text{with probability } d_k - \lfloor d_k \rfloor \end{cases},$$

where $\lfloor \cdot \rfloor$ denotes the integer part of a number.

The randomization step can be avoided, with both computational and efficiency advantages, by systematically rounding each non-integer d_k^* to the nearest integer, for instance according to a 0.5-rule for which $d_k^* = \lfloor d_k + 0.5 \rfloor$ (Chauvet, 2007; Barbiero and Mecatti 2009).

Notice that both solutions affect the characteristics of the resulting bootstrap population which might differ from the *nominal* $\mathcal{U}^*$ to an uncontrollably large extent, thus violating in the same measure the mimicking principle and the plug-in approach. For instance the constant weights $d_k^* = N/n$ guarantee a bootstrap population with the same known size of the original one, i.e. $N^* = N$, in the integer case only. Similarly the πps weights $d_k^* = \pi_k^{-1} = X/nx_k$ must be integer for both bootstrap and original populations to share the auxiliary total, i.e. $X^* = X$. Furthermore, non-integer weights usually result from any calibration procedure applied to $d_k^* = d_k = \pi_k^{-1}$ aiming at producing a bootstrap population mimicking all the known features of the original one (Barbiero *et al.*, 2012). This non-integer/rounding issue appears as worthing further investigation.

4. Simulation study

In this section we report results from a limited simulation study aimed at comparing some recent approaches to bootstrapping πps samples and to verify the equivalence illustrated in Section 3.5. The structure of the simulation is inspired by that in Antal and Tillé (2011, Section 11). In particular, a population has been considered of dimension $N = 100$ and the sample size is taken to be $n = 30$ so that the sampling fraction is particularly large.

Population values for the variable of interest are generated as $y_k = (12.5 + 3z_k^{1.2} + 15\varepsilon_k)^2 + 4000$, where $z_k \sim |N(0, 7)|$ and $\varepsilon_k \sim N(0, 1)$. The auxiliary (measure of size) variable x is generated as $x_k = y_k^{0.2}\epsilon_k$, with $\epsilon_k \sim \log N(0, 0.25)$. Figure 1 shows the pairwise scatterplots and correlation coefficients for the generated population values for the three variables y, x and z. πps sampling is conducted via Conditional Poisson using the `UPmaxentropy` function of the `sampling` package of the R environment. $M = 1000$ Monte Carlo runs and $C = 1000$ Bootstrap runs are conducted to estimate the variance of the Horvitz-Thompson estimators of four population parameters: total, Gini index and median of y, ratio of the total of y on the total of z.

Five bootstrap variance estimators have been compared. Two estimators are produced by re-sampling from a physically reconstructed bootstrap population $\mathcal{U}^*$, while the other

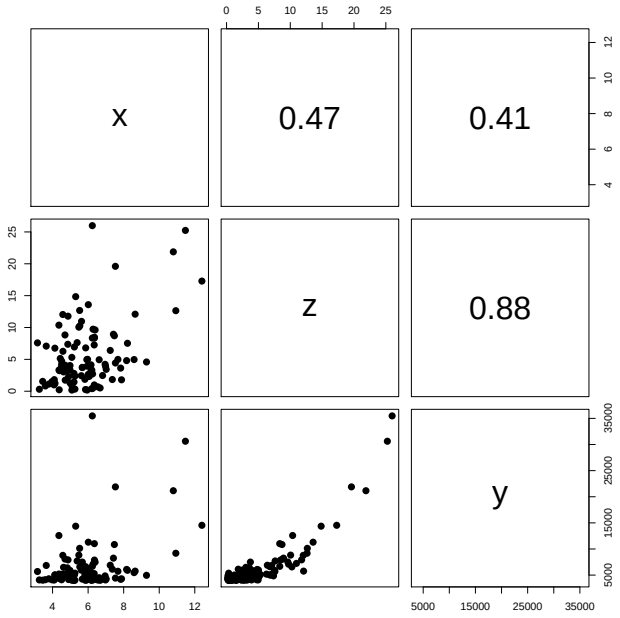

Figure 1: Pairwise scatterplots and correlation coefficients for population values of the variable of interest y, the auxiliary variable z and the size variable x generated for the simulation study.

three are derived by re-sampling directly from $\mathcal{S}$. In particular they can be classified as follows:

- Plug-in Bootstrap Population approach (re-sampling from $\mathcal{U}^*$)
  - BP-Chauvet (0.5-rule, Chauvet, 2007)
  - BP-CAL (Barbiero *et al.*, 2012), the bootstrap population is built using $d_k^* = \lfloor w_k + 0.5 \rfloor$, where w_k is a weight calibrated to match both the population size N and the auxiliary total X, namely $N^* = N$ and $X^* = X$.
- Direct bootstrap (re-sampling from $\mathcal{S}$)
  - DI-AT (Antal and Tillé, 2011, Algorithm 4)
  - Direct Plug-in Bootstrap Population approach as introduced in Section 3.5 (using the Fisher non-central Hypergeometric distribution, `biasedurn` package of the R software):
    - DI-BP-RND, using the randomization step,
    - DI-BP-Round, using the rounding approximation.

BP-Chuvet and DI-BP-Round should be equivalent ignoring random number generation variability. The following Monte Carlo measures of performance have been computed for comparison:

- Percentage Relative Bias

$$\%\text{RB} = \frac{E_{\text{MC}}[V^*(\hat{\theta}^*)] - V_{\text{MC}}(\hat{\theta})}{V_{\text{MC}}(\hat{\theta})} \cdot 100 = \frac{B}{V_{\text{MC}}(\hat{\theta})} \cdot 100;$$

Table 1: Simulation results: percentage Relative Bias, percentage Relative Root Mean Squared Error, 95% Confidence Interval coverage for the parameter based on the Normal approximation and on the Bootstrap distribution for the five estimators and the four population parameters.

	%RB	%RRMSE	Norm 95% Cov	Boot 95% Cov
TOTAL				
BP-Chauvet	-0.4	68.3	87.2	88.5
BP-Cal	-3.0	69.4	86.2	84.8
DI-AT	-1.3	72.3	86.9	88.4
DI-BP-RND	3.3	74.1	87.2	87.3
DI-BP-Round	0.5	68.1	87.2	88.7
GINI				
BP-Chauvet	-23.6	53.9	83.7	81.5
BP-Cal	-20.5	58.5	83.6	74.5
DI-AT	-31.3	55.2	82.1	82.3
DI-BP-RND	-11.0	57.6	85.1	76.6
DI-BP-Round	-16.7	54.4	84.5	79.3
MEDIAN				
BP-Chauvet	49.6	125.2	96.1	92.6
BP-Cal	35.2	111.9	96.0	92.9
DI-AT	41.9	113.7	95.4	93.3
DI-BP-RND	29.6	103.7	95.3	92.0
DI-BP-Round	38.1	114.7	95.3	91.3
RATIO				
BP-Chauvet	1.5	44.8	93.0	93.5
BP-Cal	4.6	43.9	92.9	94.1
DI-AT	3.8	44.1	93.3	93.2
DI-BP-RND	1.7	45.5	94.6	92.8
DI-BP-Round	1.1	42.3	94.1	92.8

- Percentage Relative Root Mean Squared Error

$$\%\text{RRMSE} = \sqrt{\frac{B^2 + V_{\text{MC}}[V^*(\hat{\theta}^*)]}{V_{\text{MC}}(\hat{\theta})}} \cdot 100;$$

- 95% Confidence Interval coverage based on the Normal approximation;
- 95% Confidence Interval coverage based on the Bootstrap distribution (percentile method).

Table 1 reports the results for the simulation study. The performance of the estimators is quite similar for a given parameter. There is no evidence of a uniform superiority of a method over the other in terms of efficiency. For non-linear parameters all methods appear to need improvement as far as bias is concerned (see GINI and MEDIAN). DI-BP type estimators seem to be less affected by this issue.

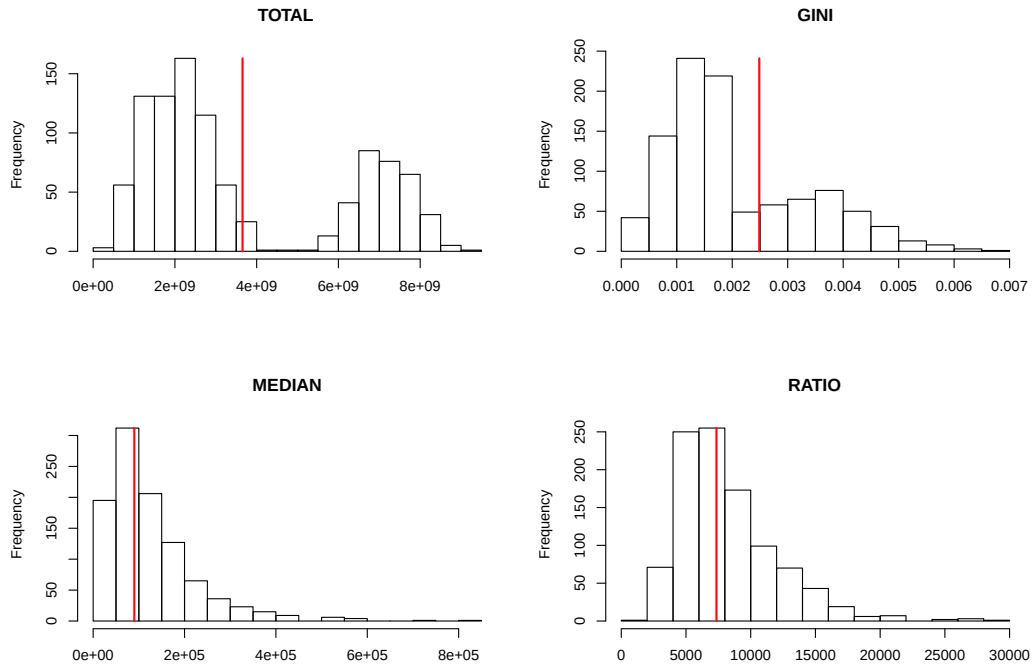

Figure 2: Monte Carlo distribution of the variance estimator DI-BP-Round for the variance of the estimators of the four parameters. The red line denotes the Monte Carlo variance.

Coverage, on the other side, is relatively better for MEDIAN and RATIO, and usually better using the Normal approximation rather than the bootstrap distribution. An explanation for this may be the following. Figure 2 shows the Monte Carlo distribution of the variance estimator DI-BP-Round for the variance of the estimators of the four parameters. The red line denotes the Monte Carlo variance over replications. The shape of the distribution is very similar for the other estimators. The first two distributions are clearly bimodal and this affects coverage. The two modes (especially when estimating TOTAL) derive for the different values the estimator takes according to whether or not a few influential points are selected in the sample. From Figure 1 it can be noted that there are about four units for which the variable of interest y takes particularly large values. The larger mode in the distribution of the estimators for TOTAL and GINI derives from those samples in which such units are selected in the sample. MEDIAN and RATIO are not affected because the former is a robust indicator and in the latter the effect of those units is mitigated by the fact that they show relatively larger values also for the auxiliary variable z. In simulations in which the generated population did not have such large values, the coverage for TOTAL and GINI is much closer to the nominal one and the shape of the distributions is clearly unimodal.

5. Conclusions

We have shown that it is possible to perform a fully plug-in approach without the need to physically reconstruct the bootstrap population for a number of popular sampling designs of increasing complexity. This provides a solution to a major limit for the application of this method otherwise appealing for respecting basic bootstrap principles. The methodology refers to the definition of a probability distribution for the re-sampling directly from the

original sample, which is proved to be equivalent to the nominal re-sampling from the bootstrap population. This can be shown to provide bootstrap unbiasedness as well second order accuracy as shown for SRS in Booth *et al.* (1994).

Moreover, the proposed methodology appears to provide a unified framework that allows to encompass other bootstrap algorithms already proposed under different approaches. See, for instance, the analogies with the naïve bootstrap (Section 3.1) and with the direct bootstrap by Antal and Tillé (2011), (Section 3.2). The rounding issue as discussed in Section 3.6 needs further attention and may be addressed by suitably modifying and generalizing the non-central multivariate Hypergeometric distribution. Finally, robustness issues as emerged from the simulation study need to be addressed to properly treat the presence of influential observations in the original sample.

REFERENCES

Antal E., and Tillé, Y. (2011), A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population, *Journal of the American Statistical Association*, 106, 534-543.

Barbiero, A., and Mecatti, F. (2009), Bootstrap algorithms for variance estimation in PS sampling. In *Complex Data Modeling and Computationally Intensive Statistical Methods*, P. Mantovan, P. Secchi, Eds., Springer-Verlag, Berlin.

Barbiero, A., Manzi, G., and Mecatti, F. (2012), Calibrated Bootstrap for probability-proportional-to-size samples, *Manuscript*.

Beaumont, J-F., and Patak, Z. (2012), On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling, *International Statistical Review*, 80-1, 127148.

Bertail, P., and Combris, P. (1997), Bootstrap geńéralisé d'un sondage. *Annales d'économie et de statistique*, 46, 4983.

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

Brewer, K.R.W., and Hanif, M. (1983), *Sampling with unequal probabilities*, Springer-Verlag, New York.

Chao M. T., and Lo A. Y. (1985), A bootstrap method for finite population, *Sankhya*, 47(A), 399-405

Chauvet, G. (2007), Méthodes de bootstrap en population finie. PhD Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Université de Rennes 2. Available at http://tel.archives-ouvertes.fr/docs/00/26/76/89/PDF/thesechauvet.pdf

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

Fog, A. (2008), Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions, *Communications in Statistics, Simulation and Computation*, 37-2, 241-257.

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

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Holmberg A. (1998), A bootstrap approach to probability proportional to size sampling. In *Proceedings of Section on Survey Research Methods*, American Statistical Association, 378-383.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, Wiley-Interscience, New York.

McCarthy, P.J., and Snowden, C.B. (1985), The bootstrap and finite population sampling. *Vital and health statistics*, Public Health Service Publication, U.S. Government Printing, Washington, DC, 95(2), 1-23

Presnell, B., and Booth, J. (1994), Resampling methods for sample surveys. *Technical report*. Available at http://www.stat.ufl.edu/ presnell/Research/TechRep/fin-pop-boot.ps

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

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

Tillé, Y. (2006), *Sampling Algorithms*, Springer-Verlag, New York.

Traat, I., Bondesson, L., Meister, K. (2004), Sampling design and sample selection through distribution theory, *Journal of Statistical Planning and Inference*, 123, 395-413.