Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Efficient unequal probability resampling from finite populations

Pier Luigi Conti [a], Fulvia Mecatti [b], Federica Nicolussi [c,*]

[a] Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le A. Moro, 5, 00185 Roma, Italy
[b] Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy
[c] Dipartimento di Economia, Management e Metodi Quantitativi, Università degli Studi di Milano, Via Festa del Perdono 7 - 20122 Milano, Italy

ARTICLE INFO

Article history:
Received 16 December 2020
Received in revised form 20 July 2021
Accepted 26 September 2021
Available online xxxx

Keywords:
Finite populations
Sampling designs
Resampling
Pseudo-population

ABSTRACT

A resampling technique for probability-proportional-to size sampling designs is proposed. It is essentially based on a special form of variable probability, without replacement sampling applied directly to the sample data, yet according to the pseudo-population approach. From a theoretical point of view, it is asymptotically correct: as both the sample size and the population size increase, under mild regularity conditions the proposed resampling design tends to coincide with the original sampling design under which sample data were collected. From a computational point of view, the proposed methodology is easy to be implemented and efficient, because it neither requires the actual construction of the pseudo-population nor any form of randomization to ensure integer weights and sizes. Empirical evidence based on a simulation study[1] indicates that the proposed resampling technique outperforms its two main competitors for confidence interval construction of various population parameters including quantiles.

1. Introduction

The use of resampling methodologies in sampling from finite populations is of considerable interest. The basic starting point consists in observing that the popular bootstrap technique, originally proposed by Efron (1979), does not work in sampling from finite populations, because of the dependence among sample units due to the sampling design. Several techniques have been proposed to overcome this problem; a nice, recent review is the paper of Mashreghi et al. (2016); cfr. also Chen et al. (2019), Rao and Wu (1988), Antal and Tillé (2011).

Among the resampling techniques for sampling designs with pre-fixed first order inclusion probabilities, πps sampling designs, for short, a special role is played by methodologies based on pseudo-populations; cfr. Mashreghi et al. (2016), Quatember (2015) for general aspects, and Conti et al. (2020) for recent theoretical contributions and a simulation study. A common feature of several resampling techniques based on pseudo-populations is their computational burden, that could be high. This motivates the study of resampling methods for πps sampling designs that share with methods based on pseudo-populations good properties in terms of variance estimation and coverage probability of confidence intervals and have a moderate computational burden. The above points are thoroughly discussed in Ranalli and Mecatti (2012), where

[*] Corresponding author.
E-mail address: federica.nicolussi@unimi.it (F. Nicolussi).

[1] R code is available as a supplementary material to the electronic version of the paper.

https://doi.org/10.1016/j.csda.2021.107366
0167-9473/© 2021 Published by Elsevier B.V.

Please cite this article as: P.L. Conti, F. Mecatti and F. Nicolussi, Efficient unequal probability resampling from finite populations, Computational Statistics and Data Analysis, https://doi.org/10.1016/j.csda.2021.107366

the problem of resampling for finite populations is addressed as a problem of sampling with replacement directly from the sample data, the *original sample* henceforth, with different drawing probabilities.

Interesting steps forward are in Quatember (2014), Quatember (2015). In particular, in Quatember (2015) an approach to resampling based on pseudo-populations with non-integer replications is proposed: Horvitz-Thompson based Boostrap, HTB, for short. HTB is still based on the physical construction of the pseudo-population, although, through the consideration of "fractions" of units, corresponding to non-integer replications of sample units, the auxiliary randomization used, for instance, in Holmberg (1998) is avoided. Essentially along the same path, in Quatember (2014) the physical construction of the pseudo-population is (implicitly) avoided. Unfortunately, as it will be seen in the sequel, Quatember's proposal does not reproduce, neither exactly nor asymptotically, as both the sample size and the population size increase, the first order inclusion probabilities of the sampling design under which sample data were collected, the *original sampling design*, henceforth. As a matter of fact, an intuitive requirement is that a resampling design, as both the sample size and the population size become large, should become closer and closer to the original sampling design. From a more formal point of view, as shown in Conti et al. (2020), the key property for the asymptotic correctness of a resampling design is that its first order inclusion probabilities should asymptotically coincide with the first order inclusion probabilities of the original sampling design. In the present paper, a new resampling technique, essentially based on sampling with replacement from the original sample, is proposed. The basic idea is to use appropriate drawing probabilities in order to reproduce, at least approximately, pre-fixed first order inclusion probabilities. Its relationships with resampling based on pseudo-populations will be discussed. The relative merits of the proposed resampling technique will be evaluated through a simulation study.

The paper is organized as follows. In Section 2, basic preliminary aspects are exposed. Section 3 deals with a general approach to resampling based on drawing of "types" from the observed sample through a *ppswor*-based technique. In Section 4 relationships with pseudo-populations are clarified; they are particularly useful to provide a sound theoretical justification of the proposed resampling technique. In Section 5 various approximations to construct drawing probabilities are exploited, and in Section 6 theoretical justifications are provided. The merits of the proposed resampling scheme are evaluated in Section 7 through a simulation study. Finally, Section 8 is devoted to conclusions.

2. Preliminary aspects and notation

Let $\mathcal{U}_N$ be a finite population of size N. A sample $\boldsymbol{s}$ is a subset of $\mathcal{U}_N$. For each unit $i \in \mathcal{U}_N$, let D_i be a Bernoulli random variable (r.v.), such that i is (is not) in the sample $\boldsymbol{s}$ whenever $D_i = 1$ ($D_i = 0$), so that $\boldsymbol{s} = \{i \in \mathcal{U}_N : D_i = 1\}$. Denote further by $\boldsymbol{D}_N$ the N-dimensional r.v. of components $D_1, \dots, D_N$. A (unordered, without replacement) sampling design P is the probability distribution of the random vector $\boldsymbol{D}_N$. From now on, the symbols E_P, V_P, C_P will denote expectation, variance and covariance w.r.t. the sampling design P. From now on, the suffix P will denote the sampling design used to select the sample $\boldsymbol{s}$.

The expectations $\pi_i = E_P[D_i]$ and $\pi_{ij} = E_P[D_i D_j]$ are the first and second order inclusion probabilities, respectively. The sample size is $n_s = D_1 + \cdots + D_N$.

In the sequel, the character of interest is denoted by $\mathcal{Y}$, and y_i is its value for unit i of the population. The population total of character $\mathcal{Y}$ is denoted by

$$t_Y = \sum_{i=1}^{N} y_i,$$

and the corresponding population mean by

$$\overline{Y}_N = N^{-1} t_Y.$$

The first order inclusion probabilities are frequently chosen to be proportional to an auxiliary variable $\mathcal{X}$. In symbols: $\pi_i \propto x_i$, where x_i is the value of $\mathcal{X}$ for unit i ($i = 1, \dots, N$). The rationale of this choice is simple: if the values of the variable of interest are positively correlated with, or, even better, approximately proportional to, the values of the auxiliary variable, then the Horvitz-Thompson estimator of the population mean will be highly efficient.

From now on, the population total of $\mathcal{X}$ and the corresponding mean will be denoted by

$$t_X = \sum_{i=1}^{N} x_i, \ \ \overline{X}_N = N^{-1} t_X,$$

respectively. With this notation, the first order inclusion probabilities are equal to:

$$\pi_i = n x_i / t_X, \ \ i = 1, \dots N. \tag{1}$$

2.1. ppswr Sampling design

Let $p_1, \dots, p_N$ be N positive numbers, with $p_1 + \cdots + p_N = 1$.

The probability proportional to size with replacement (*ppswr*, for short) sampling design of size n, with drawing probabilities $p_1, \dots, p_N$, is a sampling design where n consecutive drawings are performed. Drawings are independent, and the probability of selecting unit i at each drawing is equal to p_i. An *ordered* sample composed by units $i_1, \dots, i_n$, not necessarily distinct, has selection probability:

$$\prod_{j=1}^{n} p_{i_j}.$$

The first order inclusion probability of unit i is equal to $\pi_i = 1 - (1 - p_i)^n$. Hence, in order to have pre-fixed inclusion probabilities equal to π_is, the drawing probabilities must be equal to

$$p_i = 1 - (1 - \pi_i)^{1/n}, \quad i = 1, \dots, N. \tag{2}$$

2.2. *ppswor* Sampling design

The probability proportional to size without replacement (*ppswor*, for short) sampling design of size n, with initial drawing probabilities $p_1, \dots, p_N$ is a sampling design where n consecutive drawings are performed. The probability of selecting unit i in the final sample is proportional to p_i, and sampled units are not replaced in the population. Hence, an *ordered* sample composed by units $i_1, \dots, i_n$ has selection probability:

$$\prod_{j=1}^{n} \frac{p_{i_j}}{1 - p_{i_1} - \dots - p_{i_{j-1}}}.$$

First order inclusion probabilities for *ppswor* design are not proportional to p_is, and do not have an expression in closed form; see Hájek (1981), p. 95, where this design is termed *successive sampling*. Useful approximations are given in Hájek (1981), Rosén (1997), Rosén (2000), and described below.

Approximation R-1 cfr. (Rosén, 1997)

$$p_i \approx \log(1 - \pi_i) \Bigg/ \sum_{k=1}^{N} \log(1 - \pi_k) \ , \quad i = 1, \dots, N. \tag{3}$$

Approximation R-2 cfr. (Rosén, 2000) Let ξ_n be the (unique) root of the equation (w.r.t. t):

$$\sum_{i=1}^{N} (1 - \exp\{-p_i t\}) = n.$$

Then, the approximate relationship for inclusion probabilities

$$\pi_i \approx 1 - \exp\{-\xi_n p_i\}, \quad i = 1, \dots, N \tag{4}$$

holds. From (4), the following approximate relationship is obtained:

$$p_i \approx -\frac{1}{\xi_n} \log(1 - \pi_i), \quad i = 1, \dots, N. \tag{5}$$

Approximation H cfr. (Hájek, 1981)

$$p_i \approx \frac{\pi_i}{n} \left(1 + \frac{1}{2} \frac{n-1}{n} (\pi_i - \overline{\pi}_2), \right)$$

where

$$\overline{\pi}_2 = \frac{1}{n} \sum_{i=1}^{N} \pi_i^2. \tag{6}$$

3. Resampling for finite populations based on drawings of types from the sample

In the literature, there are several different methods for resampling from finite populations. An excellent review is in Mashreghi et al. (2016). A basic principle in finite population resampling is that the first two moments of a resampled linear statistic should match, at least approximately, the corresponding moments of the statistic w.r.t. the sample design. This principle has been first stated in Rao and Wu (1988) dubbed *scaling problem*. A detailed discussion and theoretical justifications will be given in Section 6.

In the resampling process, unit i in the original sample $\boldsymbol{s}$ will be considered as a "unit of *type* i".

As mentioned in the Introduction, here a simple principle is used: resampling a sample $\boldsymbol{s}^*$ of size n from the original sample $\boldsymbol{s}$ is essentially equivalent to draw with replacement a sample $\boldsymbol{s}^*$ of size n of types from $\boldsymbol{s}$.

This principle can be implemented in a conceptually simple way. Let $\boldsymbol{s}^* = (i_1, i_2, \dots, i_n)$ be an ordered sequence of non-necessarily distinct types in $\boldsymbol{s}$, and let $\boldsymbol{s}_j^* = (i_1, i_2, \dots, i_j)$, $j = 1, \dots, n-1$. Consider next an arbitrary array of $n \times n$ positive numbers

$$p_j^*(i\,;\, i_1, \dots, i_{j-1});\;\; i \in \boldsymbol{s},\;\; j = 1, \dots, n, \tag{7}$$

such that

$$\sum_{i \in \boldsymbol{s}} p_j^*(i\,;\, i_1, \dots, i_{j-1}) = 1,\;\; j = 1, \dots, n.$$

The probability in (7) is the probability of selecting type i at drawing j conditionally on having selected types $i_1, \dots, i_{j-1}$ in the first $j-1$ drawings. Then, the probability of selecting $\boldsymbol{s}^*$ is taken equal to:

$$p(\boldsymbol{s}^*) = p_1(i_1) p_2(i_2\,;\, i_i) \cdots p_n(i_n\,;\, i_i, \dots, i_{n-1}). \tag{8}$$

The scheme defined by (8) is completely general. To be concrete, in the sequel a special but important case will be focused, namely a sequential drawing scheme, similar to *ppswor*. Let $N_i^* \geq 1$, $i \in \boldsymbol{s}$, be the size, not necessarily integer, of type i, and let

$$N^* = \sum_{i \in \boldsymbol{s}} N_i^*$$

be the total size of all types in sample $\boldsymbol{s}$. Note that $N^* \geq n$. For each type $i \in \boldsymbol{s}$, define further an initial drawing probability p_i^*, $i \in \boldsymbol{s}$, such that

$$p_i^* > 0 \; \forall i \in \boldsymbol{s},\;\; \sum_{i \in \boldsymbol{s}} p_i^* = 1. \tag{9}$$

The *ppswor* resampling scheme consists in drawing a sample $\boldsymbol{s}^*$ of n types, not necessarily distinct, with drawing probabilities:

$$p_j(\text{Type } i | \boldsymbol{s}_{j-1}^*) = \frac{\max(0, (N_i^* - h_{i,j-1}) p_i^*)}{\sum_{l \in \boldsymbol{s}} \max(0,\, (N_l^* - h_{l,j-1}) p_l^*)},\;\; j = 1, \dots, n, \tag{10}$$

where $h_{i,j-1}$ is the number of times type i appears in $\boldsymbol{s}_{j-1}^*$. From (10) it is seen that the relationships

$$0 \leq h_{i,j} \leq N_i^* \;\; \forall i \in \boldsymbol{s},\quad \sum_{i \in \boldsymbol{s}} h_{i,j} = j \;\; \forall j = 1, \dots, n$$

hold.

As a special case, a *ppswr* resampling scheme can be obtained. Assume that $N_i^* = K^*$ for all types $i \in \boldsymbol{s}$. Then, (10) reduces to

$$p_j(\text{Type } i | \boldsymbol{s}_{j-1}^*) = \frac{\max(0, (1 - h_{i,j-1}/K^*) p_i^*)}{\sum_{l \in \boldsymbol{s}} \max(0,\, (1 - h_{l,j-1}/K^*) p_l^*)},\;\; j = 1, \dots, n,$$

and hence, by letting K^* tend to infinity and taking into account (9),

$$\lim_{K^* \to \infty} p_j(\text{Type } i | \boldsymbol{s}_{j-1}^*) = p_i^*,\;\; j = 1, \dots, n,$$

which corresponds to drawing types according to a *ppswr* scheme with drawing probabilities p_i^*s.

Of course, there are different key points to be clarified, namely the choice of the sizes N_i^*s and the choice of the initial drawing probabilities p_i^*s. They will be addressed in the subsequent Sections.

4. Relationships with pseudo-populations

The scheme of resampling types, introduced in Section 3, has clear connections with the notion of pseudo-population; cfr. Mashreghi et al. (2016), Quatember (2015), and, for large sample properties, Conti et al. (2020). A *pseudo-population* is essentially a prediction, based on sample data, of the actual population. Each unit k of the pseudo-population takes values (x_k^*, y_k^*) equal to one of the (x_i, y_i) sample pairs. Furthermore, exactly N_i^* units of the pseudo-population take the same values (x_i, y_i), with $i \in \boldsymbol{s}$; equivalently, unit i of the sample is replicated N_i^* times in the pseudo-population. More formally, a pseudo-population is represented as the set $U^* = \{(x_i, y_i, N_i^*);\ i \in \boldsymbol{s}\}$. A unit k of the pseudo-population such that $x_k^* = x_i$ and $y_k^* = y_i$ is said to be of *type* i.

If the size N_i^* of a type i, as introduced in Section 3, is integer, then it is equivalent to a pseudo-population where each sample unit i is replicated N_i^* times. This remark opens the road to different criteria for choosing N_i^*s; cfr. Conti et al. (2020).

For πps design a popular choice is the Holmberg size (Holmberg, 1998).

The Holmberg size is essentially a randomized integer-valued choice based on taking

$$N_i^* = \left\lfloor \frac{1}{\pi_i} \right\rfloor + \epsilon_i, \quad i \in \boldsymbol{s},$$

where $\lfloor\ \rfloor$ denotes the integer part (floor) and ϵ_is are independent Bernoulli r.v.s. with

$$P(\epsilon_i = u|\boldsymbol{s}) = r_i^u (1 - r_i)^{1-u}, \quad u \in \{0, 1\},\ i \in \boldsymbol{s}$$

with

$$r_i = \frac{1}{\pi_i} - \left\lfloor \frac{1}{\pi_i} \right\rfloor, \quad i \in \boldsymbol{s}.$$

Notice that integer-valued N_i^*s are mandatory in order to actually build up the pseudo-population. According to the principle of resampling types illustrated in the previous Section, such a request may be relaxed by removing the additional uncertainty due to the randomization.

The Horvitz-Thompson size is essentially a non-randomized version of the Holmberg size; it is based on taking:

$$N_i^* = \frac{1}{\pi_i} = \frac{t_X}{n x_i}, \quad i \in \boldsymbol{s}.$$

The values N_i^*s are not necessarily integer, which is often the case in practice. Moreover, the Horvitz-Thompson size has the important property

$$\sum_{i \in \boldsymbol{s}} N_i^* x_i = \sum_{i=1}^{N} x_i,$$

namely it is calibrated w.r.t. the total of the auxiliary variable X.

The combination of the proposed *ppswor* resampling of types and the HT size allows an asymptotically correct resampling based on a pseudo-population without requiring neither its actual construction, nor the constraint of integer sizes. As a consequence, both computational and precision advantages can be expected.

5. Drawing probabilities for resampling

The drawing probabilities p_i^*s used in resampling should be chosen in order to ensure, at least approximately, inclusion probabilities proportional to x_is. In this way, the resampling scheme becomes asymptotically correct. Hence, the target first order inclusion probability of unit k of type i of the pseudo population is

$$\begin{aligned}\pi_k^* &= \pi_{(i)}^* \\ &= n x_k^* / \sum_{k=1}^{N^*} x_k^* \\ &= n x_i / \sum_{i \in \boldsymbol{s}} x_i N_i^* \\ &= n x_i / t_X^*. \end{aligned} \tag{11}$$

As a consequence of Rosén (1997), Rosén (2000), if both the population size N and the sample size n increase, the first order inclusion probabilities of the corresponding resampling scheme are asymptotically linear in x_is, and then asymptotically equivalent to the first order inclusion probabilities of the original sampling design. In the sequel, various approximations for p_i^*s, based on those listed in Section 2.2, are examined.

1. Approximation R-1

Using the notation introduced in (3) and based on (11), the relationship

$$p_{i,R1}^* \approx \log\left(1-\frac{nx_i}{t_X^*}\right) \Big/ \sum_{l\in\mathbf{s}} N_l^* \log\left(1-\frac{nx_l}{t_X^*}\right) \tag{12}$$

holds for all the N_i^* pseudo-population units of type i.

2. Approximation R-2

A second solution, computationally heavier than R-1, can be based on approximation R-2 (5). The major difficulty is that the term ξ_n^*, which is the (unique) solution of the equation

$$\sum_{i\in\mathbf{s}} N_i^*\left(1-\exp\left\{-p_i^* t\right\}\right) = n$$

cannot be directly computed on the basis of target first order inclusion probabilities. To this purpose, the following iterative algorithm can be used.

0. Set $m=0$, $\pi_{(i)}^*(m)=\pi_{(i)}^*$, $i\in\mathbf{s}$, and take a (small) threshold $\delta>0$. Go to Step 1.
1. Compute

$$p_i^*(m) = \log\left(1-\pi_{(i)}^*(m)\right) \Big/ \sum_{l\in\mathbf{s}} N_l^* \log\left(1-\pi_{(l)}^*(m)\right), \quad i\in\mathbf{s}.$$

   Go to Step 2.
2. Compute $\xi_n^*(m)$ as the solution of the equation:

$$\sum_{i\in\mathbf{s}} N_i^*\left(1-\exp\left\{-p_i^*(m)t\right\}\right) = n.$$

   Go to Step 3.
3. Compute

$$\pi_i^*(m+1) = 1-\exp\left\{-\xi_n^*(m)p_i^*(m)\right\}, \quad i\in\mathbf{s}. \tag{13}$$

   Go to Step 4.
4. Set $m\to m+1$. If $|\pi_i^*(m+1)-\pi_i^*|<\delta$ for every $i\in\mathbf{s}$, then go to Step 5. Otherwise, go to Step 1.
5. Stop. Set

$$p_{i,R2}^* = p_i^*(m), \quad i\in\mathbf{s}. \tag{14}$$

The functions $f_m=\sum_{i\in\mathbf{s}} N_i^*\left(1-e^{-p_i^*(m)t}\right)$ are concave, twice differentiable, and satisfy the Lipschitz condition w.r.t. t for each fixed set of $p_i^*(m)$s. Furthermore, viewed as a function of $p_i^*(m)$s, f_m is Lipschitz w.r.t. $p_i^*(m)$, $i\in\mathbf{s}$, for each fixed t on each compact subset $[0,K]$ of the non-negative real half-line. As a consequence, if the starting values $p_i^*(0)$ are close to p_i^*s, the same holds $p_i^*(m)$, $m\geq 1$, generated at each iteration. In addition, the same reasoning holds for $\xi_n^*(m)$s. More formally, cfr. Kelley (1995), the sequences $(p_i^*(m);\ m\geq 1)$, $(\xi_i^*(m);\ m\geq 1)$ are Picard sequences, and hence convergent. This is a local property, in the sense that it holds true provided that the initial values $p_i^*(0)$s are sufficiently close to p_i^*s. It is suggested to set the initial values $p_i^*(0)$s *via* R-1 approximation, so that they are in principle close to p_i^*s. The simulation study in Section 7 empirically confirms the convergence of the algorithm.

3. Approximation H

The starting point consists in using (11), which implies

$$\begin{aligned}\overline{\pi}_2^* &= \frac{1}{n}\sum_k \pi_k^{*2}\\ &= \frac{1}{n}\sum_{i\in\mathbf{s}} N_i^*\left(\frac{nx_i}{t_X^*}\right)^2.\end{aligned}$$

Hence, the drawing probabilities that approximate the target inclusion probabilities (11) are equal to

$$p^*_{i,H} = \frac{x_i}{t^*_X}\left\{1+\frac{1}{2}\frac{n-1}{n}\left(\frac{nx_i}{t^*_X}-\overline{\pi}^*_2\right)\right\}, \tag{15}$$

for all N_i^* units of type i, with

$$\sum_{i\in\boldsymbol{s}} N_i^* p^*_{i,H} = 1.$$

6. Theoretical justifications

The goal of the present section is to provide a few theoretical justifications to the resampling scheme developed so far. As shown in Conti et al. (2020), if the sampling design possesses asymptotically maximal entropy, and if N_i^*s satisfy appropriate regularity conditions (the most important one being that their expectations are asymptotically equivalent to π_i^{-1}s), then the resampling design based on (normalized) Conditional Poisson design, also known as Maximum Entropy design or rejection sampling, is fully justified from an asymptotic viewpoint. As a consequence, it is also justified on the basis of Rao's *scaling problem* mentioned in Section 3, namely the principle of matching the first two moments of linear statistics. In the sequel, the first and second order inclusion probabilities for the normalized Conditional Poisson design will be denoted by $\pi^{*R}_{(i)}$, $\pi^{*R}_{(ij)}$ for all pairs of distinct units in the pseudo population U^*, of type i and type j, respectively, with $j \neq i$. Of course, $\pi^{*R}_{(i)}$ is equal to nx_i/t^*_X for all units of type i.

Resampling design based on *ppswor* does not possess the same asymptotic justification, although it possesses good asymptotic properties: cfr. Rosén (1972). Moreover, it possesses good properties in terms of Rao's principle cited above, as it will be now illustrated.

Denote by $\pi^{*S}_{(i)}$, $\pi^{*S}_{(ij)}$ the first and second order inclusion probabilities for units of type i, $j \neq i$, let $f^*_N = n/N^*$ be the resampling fraction, and $\overline{X}^* = t^*_X/N^*$. Note that the target first order inclusion probabilities (11) are also equal to

$$\pi^*_{(i)} = f^*_N \frac{x_i}{\overline{X}^*}. \tag{16}$$

When approximation R-1 (or R-2) is used:

$$\begin{aligned} p^*_{(i)} &= \log\left(1-\pi^{*R}_{(i)}\right) / \sum_{i\in\boldsymbol{s}} N_i^* \log\left(1-\pi^{*R}_{(i)}\right) \\ &= \log\left(1-f^*_N\frac{x_i}{\overline{X}^*}\right) / \sum_{i\in\boldsymbol{s}} N_i^* \log\left(1-f^*_N\frac{x_i}{\overline{X}^*}\right), \end{aligned} \tag{17}$$

then the resampling design based on *ppswor* possesses not only first order inclusion probabilities proportional to x_is, $\pi^{*S}_{(i)} \simeq nx_i/t^*_X$, but also second order inclusion probabilities that are close to $\pi^{*R}_{(ij)}$s. As a consequence, the proposed resampling based on *ppswor* it is fully justified on the basis of Rao principle for the first moment of linear statistics, and approximately justified for the second moment of linear statistics. This point is clarified in the subsequent Proposition 1.

Define first

$$\Delta^{*R}_{(ij)} = \pi^{*R}_{(ij)} - \pi^{*R}_{(i)}\pi^{*R}_{(i)}, \quad \Delta^{*S}_{(ij)} = \pi^{*S}_{(ij)} - \pi^{*S}_{(i)}\pi^{*S}_{(i)}, \quad i \neq j.$$

As a consequence of (1.9), (1.10) in Hájek (1973), and taking into account that, up to a term asymptotically negligible,

$$\pi^{*S}_i = \pi^{*R}_i = n\frac{x_i}{t^*_X}, \tag{18}$$

hence

$$\Delta^{*R}_{ij} \sim \frac{\pi^{*R}_i(1-\pi^{*R}_i)\pi^{*R}_j(1-\pi^{*R}_j)}{\sum_{i\in\boldsymbol{s}} N_i^*\pi^{*R}_i(1-\pi^{*R}_i)} \quad i \neq j \in \boldsymbol{s}, \tag{19}$$

and, in view of (18),

$$\begin{aligned} \Delta^{*S}_{ij} &\sim \frac{\pi^{*S}_i(1-\pi^{*S}_i)\pi^{*S}_j(1-\pi^{*S}_j)}{\sum_{i\in\boldsymbol{s}} N_i^*\pi^{*S}_i(1-\pi^{*S}_i)}\left\{1-\left(1-\frac{\overline{\pi}^{*S}p^*_i}{\overline{p}^*\pi^{*S}_i}\right)\left(1-\frac{\overline{\pi}^{*S}p^*_j}{\overline{p}^*\pi^{*S}_j}\right)\right\} \\ &= \Delta^{*R}_{ij}\left\{1-\left(1-\frac{\overline{\pi}^*p^*_i}{\overline{p}^*\pi^{*R}_i}\right)\left(1-\frac{\overline{\pi}^*p^*_j}{\overline{p}^*\pi^{*R}_j}\right)\right\}, \end{aligned} \tag{20}$$

where $\sim$ means that the ratio of both sides converges to 1 as both the sample size and the population size increase, and

$$\overline{\pi}^* = \sum_{i\in \boldsymbol{s}} N_i^* \pi_i^{*S}(1-\pi_i^{*S}) = \sum_{i\in \boldsymbol{s}} N_i^* f_N^* \frac{x_i}{\overline{X}^*}\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right), \tag{21}$$

$$\begin{aligned}\overline{p}^* &= \sum_{i\in \boldsymbol{s}} N_i^* p_i^*(1-\pi_i^{*S}) \\ &= \sum_{i\in \boldsymbol{s}} N_i^* \left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right)\log\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right) \Big/ \sum_{i\in \boldsymbol{s}} N_i^* \log\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right).\end{aligned} \tag{22}$$

Proposition 1. *Consider the normalized Conditional Poisson resampling design with* $\pi_{(i)}^* = f_N^* x_i/\overline{X}^*$, *and* ppswor *resampling design with drawing probabilities* p_i^* *proportional to* $\log\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right)$. *Then:*

$$\frac{\Delta_{ij}^{*R} - \Delta_{ij}^{*S}}{\Delta_{ij}^{*R}} = O(f_N^{*2}). \tag{23}$$

Result (23) is interesting essentially for one reason. The resampling variance of linear statistics depends on the terms $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. If the sampling design possesses high entropy, then the normalized Conditional Poisson resampling design is asymptotically correct; as a consequence, the resampling variance of linear statistics is asymptotically equivalent to their sampling variance. Proposition 1 tells us that, up to a term $O\left(f_N^{*2}\right)$, the same holds for *ppswor* resampling design, provided that the corresponding first-order inclusion probabilities are, at least asymptotically, equal to $f_N^* x_i/\overline{X}^*$. Since the *square* of the resampling fraction, f_N^{*2}, is usually very small, on one hand *ppswor* resampling design possesses properties that are close to that of normalized Conditional Poisson resampling design. On the other hand, it offers a considerable computational advantage. This remark is made it stronger by a simple consideration: under mild regularity conditions, N^*/N tends in probability to 1, and hence the resampling fraction f_N^* tends to be asymptotically equivalent, in probability, to the sampling fraction $f_N = n/N$.

7. Simulation study

In order to test the empirical performance of the proposed *ppswor* resampling of types procedure, as illustrated in Section 3, a simulation exercise has been conducted, based on the Horvitz-Thompson size pseudo-population (see Section 4) and under each of the three alternative options described in Section 5 to approximate the resampling (drawing) probabilities. Three further bootstrap methods available in the literature have been considered in the present simulation study as main competitors of our proposal, namely:

1) Quatember's algorithm (Quatember, 2014), dubbed Q, that is comparable both in terms of being based on a pseudo-population and, at the same time, being simplified by resampling directly from the (original) sample under a *ppswor* design, as mentioned in Section 1;
2) Holmberg's method (Holmberg, 1998), dubbed *Holm*, that involves a resampling based on a pseudo-population under a randomized version of the Horvitz-Thompson's size, see Section 4. However Holmberg's method requires the actual construction of the pseudo-population and then to re-sample in it by mimicking the original sampling design. Thus, in a sense, it acts as a benchmark w.r.t. computational efficiency of our proposed bootstrap algorithm; and
3) HTB technique (Quatember, 2015), illustrated in Section 1 that, similarly to the Holmberg's method above, requires the actual construction of a pseudo-population although by allowing non-integer replications of sample units.

Six scenarios composed by populations of increasing size N from 200 to 5,000 are explored. For the sake of comparability, they are generated as in Antal and Tillé (2011), Conti et al. (2020). In detail, the study variable $\mathcal{Y}$ has been generated according to the model $y_i = \left(12.5 + 3w_i^{1.2} + \sigma\epsilon_i\right)^2 + 4,000$ where $w_i \sim |N(0,7)|$, $\epsilon_i \sim N(0,1)$ and $\sigma = 15$. Selection probabilities are taken proportional to values x_i generated from $\mathcal{X} = \mathcal{Y}^{0.2} \cdot LogN(0, 0.025)$ where $LogN$ denotes a Log-normal probability distribution. The choice of a relation of approximate proportionality between the study variable and the auxiliary variable leads to selection probabilities approximately proportional to $\mathcal{Y}$, which is the rationale behind a πps sampling. The correlation coefficient between $\mathcal{Y}$ and $\mathcal{X}$ is fairly moderate, about 0.8 uniformly across all simulated scenarios. For each population, 1,000 samples are simulated under a Pareto sampling design. The latter choice has two main motivations. First of all, Pareto sampling design is very simple to be implemented and computationally inexpensive. In the second place, it possesses good properties because of its high entropy, and because it is heuristically considered as almost equivalent to the asymptotically maximum entropy Rao-Sampford design (Bondesson et al., 2006). Two sampling fractions n/N have been used, namely 0.04 and 0.20, with the twofold aim of evaluating small to large finite sample sizes and to enhance the simulation of the Hájek asymptotic setup, see (Hájek, 1981, Chapter 3).

Table 1
Simulated scenarios.

Scenarios						
Population size N	200	400	800	1200	2400	5,000
Sampling fraction n/N	0.04 **0.20**					
Sample size n	8 **40**	16 **80**	32 **160**	48 **240**	96 **480**	200 **1,000**

The estimation of three population parameters is investigated:

- *Population mean* $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$;
- *Population median* $Me_Y = Me_Y = \inf\{y : F_N(y) \geq 0.5\}$;
- *Population third quartile* $Q3_Y = \inf\{y : F_N(y) \geq 0.75\}$

where $F_N(y) = N^{-1} \sum_{i=1}^{N} I_{(y_i \leq y)}$ is the population distribution function, $I_{(y_i \leq y)}$ being equal to 1 whenever $y_i \leq y$, and 0 otherwise. As estimators of the above parameters, the Hájek estimators

- $\hat{\bar{Y}}_H = \sum_{i=1}^{N} D_i \pi_i^{-1} y_i / \sum_{i=1}^{N} D_i \pi_i^{-1}$;
- $\hat{Me}_Y = \inf\{y : \hat{F}_H(y) \geq 0.5\}, \quad y \in \mathbb{R}$;
- $\hat{Q}3_Y = \inf\{y : \hat{F}_H(y) \geq 0.75\}, \quad y \in \mathbb{R}$

have been considered, where $\hat{F}_H(y) = \sum_{i=1}^{N} D_i \pi_i^{-1} I_{(y_i \leq y)} / \sum_{i=1}^{N} D_i \pi_i^{-1}$ is the Hájek estimator of $F_N(y)$. In addition, for the population mean also the Horvitz-Thompson (HT) estimator is considered

$$\hat{\bar{Y}}_{HT} = N^{-1} \sum_{i=1}^{N} D_i \pi_i^{-1} y_i.$$

HT estimator is popular in practice because of its unbiasedness, although it is frequently less efficient than the asymptotically unbiased Hájek estimator.

The simulated scenarios, are summarized in Table 1.

For each simulated sample, 1,000 bootstrap runs are performed under the six resampling methods mentioned above and dubbed R-1, R-2, H, Q, Holm and HTB respectively.

The methods are compared in terms of both Empirical Coverage (EC) and Average Length (AL) of resampling-based Confidence Interval (CI). Two popular bootstrap methods have been used: 1) the bootstrap-percentile method, i.e. by the direct use of the quantiles of the bootstrap replicates; and 2) the method based on the standard Normal quantiles coupled with the bootstrap estimate of the standard error, dubbed bootstrap-stdN.

Complete simulation results are reported in Tables 2–5. As additional information that complement EC and AL, simulated relative bias (RB) and relative root mean squared error (RRMSE) of the point bootstrap variance estimates, are reported in Appendix B, Table 6.

As a general remark, for small sample fraction and small sizes all resampling simulated tend to perform similarly for all estimators and for both types of bootstrap CIs. In particular it is noticeable how simulation results for HTB are quite uniformly close to Holm results. This effect is also noted in (Quatember, 2015, p.82) and it is somehow expected as a consequence of the similarity between the two pseudo-populations actually constructed under the two methods, that in fact differ at most for a subset of n units. However, such differences also reflect on the re-sampling inclusion probabilities, so that the HTB pseudo-population includes a fraction n/N of units whose π_i^*, with probability 1, are not proportional to x_i asymptotically. This latter can be the reason for the relatively growing differences between HTB and Holm as population and sample sizes increase.

It is also noticeable the superiority of H against HT for estimating the population mean. This is apparent for the bootstrap-stdN CIs in Table 4, where H estimation systematically provides better EC, and it is also shown by the bootstrap-percentile CIs in Table 2 where, as the population size increases, HT estimation tends to produce too conservative CIs and larger ALs.

For the higher sampling fraction and sizes, namely when the Hájek asymptotic setup is simulated more effectively, differences are more evident.

Focusing on H estimation and large sample sizes (Tables 2, 4), simulation results reveal some general pattern. Our new resampling method is associated with the best results, quite uniformly for the three estimated parameters and both types of CIs. Such empirical evidence is consistent with the theoretical properties illustrated in Section 6. Our proposed bootstrap algorithm seems to be able to improve upon all simulated competitors. In particular it provides CIs with smaller AL than HTB method for comparable ECs. Our proposed resampling scheme also gives results better than Quatember's method (Q), which seems likely to bear the worst ECs and a tendency to shrinking ALs as N increases. This effect is more enhanced in case of bootstrap-percentile CIs while less evident for bootstrap-stdN CIs. Finally, the proposed bootstrap based on directly

Table 2
Empirical Coverage (EC) and Average Length (AL) of CIs bootstrap-percentile method for the population mean, for increasing population sizes, two sample fractions 4% **20%** and two point estimates Horvitz-Thompson and Hájek.

N	200		400		800		1200		2400		5000	
	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
$\hat{\bar{Y}}_{HT}$												
HTB	0.78	0.61	0.84	0.60	0.92	0.47	0.92	0.44	0.93	0.35	0.96	0.26
	0.92	**0.59**	**0.92**	**0.49**	**0.89**	**0.37**	**0.93**	**0.29**	**0.91**	**0.23**	**0.84**	**0.18**
Holm	0.74	0.53	0.84	0.51	0.91	0.43	0.93	0.40	0.93	0.32	0.97	0.24
	0.9	**0.41**	**0.95**	**0.31**	**0.97**	**0.26**	**0.97**	**0.22**	**0.98**	**0.17**	**0.99**	**0.13**
Q	0.74	0.52	0.84	0.50	0.91	0.42	0.92	0.40	0.93	0.32	0.97	0.24
	0.93	**0.44**	**0.97**	**0.33**	**0.99**	**0.28**	**0.99**	**0.24**	**0.9**	**0.18**	**0.98**	**0.14**
R-1	0.74	0.53	0.83	0.50	0.91	0.42	0.92	0.40	0.93	0.31	0.97	0.24
	0.9	**0.40**	**0.94**	**0.31**	**0.97**	**0.26**	**0.97**	**0.21**	**0.98**	**0.17**	**0.99**	**0.13**
R-2	0.74	0.53	0.83	0.50	0.91	0.42	0.92	0.40	0.93	0.31	0.97	0.24
	0.9	**0.40**	**0.94**	**0.31**	**0.97**	**0.26**	**0.97**	**0.22**	**0.98**	**0.17**	**0.99**	**0.13**
H	0.74	0.53	0.83	0.50	0.91	0.42	0.92	0.40	0.93	0.31	0.97	0.24
	0.91	**0.41**	**0.95**	**0.31**	**0.98**	**0.26**	**0.98**	**0.22**	**0.99**	**0.17**	**0.99**	**0.13**
$\hat{\bar{Y}}_{H}$												
HTB	0.86	0.56	0.89	0.42	0.9	0.29	0.91	0.25	0.89	0.19	0.95	0.14
	0.88	**0.26**	**0.91**	**0.22**	**0.91**	**0.15**	**0.86**	**0.13**	**0.81**	**0.10**	**0.82**	**0.08**
Holm	0.86	0.55	0.89	0.40	0.9	0.28	0.91	0.24	0.9	0.18	0.93	0.13
	0.87	**0.23**	**0.9**	**0.17**	**0.92**	**0.13**	**0.91**	**0.11**	**0.92**	**0.08**	**0.93**	**0.06**
Q	0.86	0.56	0.88	0.40	0.9	0.28	0.91	0.24	0.89	0.18	0.92	0.13
	0.84	**0.22**	**0.81**	**0.16**	**0.77**	**0.12**	**0.71**	**0.10**	**0.62**	**0.08**	**0.46**	**0.06**
R-1	0.85	0.55	0.89	0.41	0.9	0.28	0.9	0.24	0.9	0.18	0.92	0.13
	0.88	**0.24**	**0.91**	**0.17**	**0.93**	**0.13**	**0.92**	**0.11**	**0.93**	**0.09**	**0.95**	**0.06**
R-2	0.85	0.55	0.89	0.41	0.9	0.29	0.9	0.24	0.9	0.18	0.93	0.13
	0.88	**0.24**	**0.91**	**0.17**	**0.93**	**0.13**	**0.92**	**0.11**	**0.93**	**0.08**	**0.95**	**0.06**
H	0.86	0.55	0.89	0.41	0.9	0.28	0.9	0.24	0.89	0.18	0.93	0.13
	0.88	**0.23**	**0.9**	**0.17**	**0.91**	**0.13**	**0.88**	**0.11**	**0.88**	**0.08**	**0.88**	**0.06**

Table 3
Empirical Coverage (EC) and Average Length (AL) of CIs bootstrap-percentile method for the population median (Me) and 0.75 percentile (Q3), for increasing population sizes, and two sample fractions 4% **20%**.

N	200		400		800		1200		2400		5000	
	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
$\hat{Me}_Y$												
HTB	0.77	0.80	0.78	0.57	0.79	0.38	0.73	0.34	0.80	0.25	0.88	0.20
	0.71	**0.42**	**0.78**	**0.29**	**0.78**	**0.21**	**0.77**	**0.20**	**0.72**	**0.15**	**0.71**	**0.12**
Holm	0.76	0.79	0.77	0.44	0.79	0.36	0.78	0.33	0.83	0.24	0.86	0.19
	0.78	**0.34**	**0.83**	**0.23**	**0.83**	**0.17**	**0.85**	**0.16**	**0.87**	**0.12**	**0.92**	**0.09**
Q	0.78	0.83	0.78	0.55	0.79	0.36	0.77	0.33	0.81	0.24	0.85	0.18
	0.73	**0.30**	**0.74**	**0.19**	**0.75**	**0.15**	**0.76**	**0.13**	**0.7**	**0.09**	**0.69**	**0.06**
R-1	0.75	0.79	0.79	0.57	0.8	0.37	0.78	0.33	0.82	0.24	0.88	0.19
	0.79	**0.35**	**0.84**	**0.23**	**0.84**	**0.18**	**0.85**	**0.16**	**0.88**	**0.12**	**0.92**	**0.09**
R-2	0.75	0.79	0.79	0.57	0.8	0.37	0.78	0.34	0.82	0.24	0.87	0.19
	0.79	**0.35**	**0.84**	**0.23**	**0.84**	**0.18**	**0.85**	**0.16**	**0.87**	**0.12**	**0.91**	**0.09**
H	0.75	0.78	0.79	0.56	0.8	0.37	0.79	0.33	0.82	0.24	0.87	0.19
	0.78	**0.34**	**0.82**	**0.22**	**0.82**	**0.17**	**0.83**	**0.16**	**0.85**	**0.11**	**0.88**	**0.09**
$\hat{Q}3_Y$												
HTB	0.78	0.03	0.78	0.70	0.82	0.50	0.85	0.41	0.86	0.33	0.85	0.22
	0.76	**0.49**	**0.75**	**0.43**	**0.75**	**0.32**	**0.74**	**0.25**	**0.81**	**0.21**	**0.64**	**0.15**
Holm	0.77	0.90	0.77	0.66	0.81	0.47	0.86	0.39	0.86	0.30	0.87	0.20
	0.85	**0.38**	**0.82**	**0.29**	**0.82**	**0.23**	**0.82**	**0.18**	**0.92**	**0.13**	**0.89**	**0.10**
Q	0.79	0.94	0.77	0.67	0.81	0.47	0.86	0.38	0.86	0.29	0.85	0.19
	0.75	**0.32**	**0.73**	**0.23**	**0.72**	**0.17**	**0.68**	**0.13**	**0.66**	**0.08**	**0.52**	**0.06**
R-1	0.77	0.90	0.78	0.68	0.81	0.48	0.88	0.40	0.86	0.30	0.87	0.20
	0.86	**0.39**	**0.84**	**0.29**	**0.83**	**0.23**	**0.83**	**0.18**	**0.93**	**0.14**	**0.91**	**0.10**
R-2	0.72	0.90	0.55	0.68	0.41	0.48	0.35	0.40	0.26	0.30	0.18	0.20
	0.86	**0.39**	**0.84**	**0.29**	**0.83**	**0.23**	**0.83**	**0.18**	**0.93**	**0.14**	**0.91**	**0.10**
H	0.77	0.90	0.78	0.68	0.81	0.48	0.88	0.40	0.86	0.30	0.87	0.20
	0.82	**0.37**	**0.8**	**0.27**	**0.8**	**0.21**	**0.78**	**0.17**	**0.89**	**0.12**	**0.84**	**0.09**

Table 4
Empirical Coverage (EC) and Average Length (AL) of CIs bootstrap-stdN method for the population mean, for increasing population sizes, two sample fractions 4% **20%** and two point estimates Horvitz-Thompson and Hájek.

N	200		400		800		1200		2400		5000	
	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
$\hat{\bar{Y}}_{HT}$												
HTB	0.65	0.74	0.74	0.72	0.78	0.65	0.79	0.54	0.81	0.41	0.85	0.30
	0.77	**0.73**	**0.8**	**0.65**	**0.85**	**0.43**	**0.85**	**0.34**	**0.87**	**0.31**	**0.9**	**0.21**
Holm	0.66	0.56	0.73	0.53	0.77	0.44	0.8	0.42	0.82	0.33	0.86	0.25
	0.77	**0.43**	**0.8**	**0.32**	**0.84**	**0.27**	**0.83**	**0.22**	**0.87**	**0.17**	**0.91**	**0.13**
Q	0.65	0.54	0.73	0.52	0.78	0.44	0.8	0.42	0.82	0.33	0.87	0.25
	0.79	**0.45**	**0.82**	**0.34**	**0.86**	**0.29**	**0.85**	**0.24**	**0.89**	**0.19**	**0.92**	**0.15**
R-1	0.65	0.55	0.73	0.52	0.77	0.43	0.79	0.42	0.81	0.32	0.86	0.24
	0.77	**0.42**	**0.8**	**0.32**	**0.84**	**0.27**	**0.83**	**0.22**	**0.87**	**0.17**	**0.91**	**0.13**
R-2	0.65	0.55	0.73	0.52	0.77	0.43	0.79	0.42	0.81	0.32	0.86	0.24
	0.77	**0.42**	**0.8**	**0.32**	**0.84**	**0.27**	**0.83**	**0.22**	**0.87**	**0.17**	**0.91**	**0.13**
H	0.65	0.55	0.73	0.52	0.77	0.44	0.8	0.42	0.81	0.32	0.86	0.24
	0.77	**0.43**	**0.81**	**0.32**	**0.85**	**0.27**	**0.83**	**0.23**	**0.87**	**0.17**	**0.23**	**0.14**
$\hat{\bar{Y}}_{H}$												
HTB	0.9	0.57	0.9	0.43	0.92	0.30	0.93	0.25	0.92	0.20	0.93	0.14
	0.91	**0.26**	**0.9**	**0.23**	**0.91**	**0.16**	**0.88**	**0.13**	**0.9**	**0.11**	**0.92**	**0.08**
Holm	0.89	0.57	0.91	0.41	0.92	0.29	0.93	0.24	0.92	0.19	0.93	0.13
	0.9	**0.24**	**0.91**	**0.17**	**0.93**	**0.13**	**0.91**	**0.11**	**0.91**	**0.08**	**0.94**	**0.06**
Q	0.89	0.58	0.91	0.41	0.92	0.29	0.93	0.24	0.91	0.18	0.93	0.13
	0.89	**0.22**	**0.9**	**0.16**	**0.92**	**0.12**	**0.89**	**0.10**	**0.9**	**0.08**	**0.92**	**0.06**
R-1	0.89	0.56	0.92	0.42	0.92	0.29	0.94	0.25	0.91	0.18	0.94	0.13
	0.91	**0.24**	**0.93**	**0.17**	**0.93**	**0.13**	**0.92**	**0.11**	**0.92**	**0.09**	**0.94**	**0.06**
R-2	0.89	0.56	0.92	0.42	0.92	0.29	0.94	0.25	0.91	0.18	0.94	0.13
	0.91	**0.24**	**0.93**	**0.17**	**0.93**	**0.13**	**0.92**	**0.11**	**0.92**	**0.09**	**0.94**	**0.06**
H	0.89	0.56	0.92	0.42	0.92	0.29	0.94	0.25	0.91	0.18	0.93	0.13
	0.9	**0.24**	**0.92**	**0.17**	**0.93**	**0.13**	**0.91**	**0.11**	**0.91**	**0.08**	**0.94**	**0.06**

Table 5
Empirical Coverage (EC) and Average Length (AL) of CIs bootstrap-stdN method for the population median (Me) and 0.75 percentile (Q3), for increasing population sizes, and two sample fractions 4% **20%**.

N	200		400		800		1200		2400		5000	
	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL	EC	AL
$\hat{Me}_Y$												
HTB	0.93	0.88	0.93	0.61	0.91	0.40	0.9	0.35	0.91	0.26	0.92	0.20
	0.9	**0.41**	**0.90**	**0.31**	**0.9**	**0.20**	**0.91**	**0.19**	**0.89**	**0.16**	**0.91**	**0.11**
Holm	0. 92	0.86	0.92	0.57	0.91	0.38	0.89	0.34	0.91	0.25	0.93	0.19
	0.9	**0.36**	**0.92**	**0.24**	**0.9**	**0.18**	**0.9**	**0.16**	**0.92**	**0.12**	**0.93**	**0.09**
Q	0.92	0.89	0.92	0.58	0.91	0.37	0.89	0.34	0.91	0.24	0.92	0.19
	0.88	**0.33**	**0.89**	**0.22**	**0.88**	**0.17**	**0.88**	**0.15**	**0.88**	**0.11**	**0.9**	**0.08**
R1	0.92	0.86	0.92	0.59	0.92	0.38	0.90	0.34	0.91	0.24	0.93	0.19
	0.9	**0.36**	**0.91**	**0.24**	**0.91**	**0.18**	**0.9**	**0.16**	**0.92**	**0.12**	**0.94**	**0.09**
R2	0.92	0.86	0.92	0.59	0.92	0.38	0.9	0.34	0.91	0.24	0.92	0.19
	0.9	**0.36**	**0.91**	**0.24**	**0.91**	**0.18**	**0.9**	**0.16**	**0.92**	**0.12**	**0.93**	**0.09**
H	0.91	0.86	0.92	0.59	0.91	0.38	0.9	0.34	0.91	0.24	0.92	0.19
	0.9	**0.35**	**0.91**	**0.24**	**0.9**	**0.18**	**0.89**	**0.16**	**0.92**	**0.12**	**0.93**	**0.09**
$\hat{Q}3_Y$												
HTB	0.93	1.00	0.9	0.76	0.93	0.53	0.93	0.43	0.92	0.33	0.91	0.22
	0.93	**0.47**	**0.91**	**0.43**	**0.88**	**0.29**	**0.9**	**0.22**	**0.92**	**0.19**	**0.9**	**0.12**
Holm	0.93	0.98	0.91	0.70	0.92	0.49	0.93	0.40	0.91	0.30	0.92	0.21
	0.93	**0.40**	**0.93**	**0.30**	**0.9**	**0.23**	**0.9**	**0.18**	**0.94**	**0.13**	**0.92**	**0.10**
Q	0.93	1.01	0.9	0.70	0.92	0.49	0.93	0.40	0.91	0.30	0.92	0.20
	0.91	**0.38**	**0.90**	**0.28**	**0.87**	**0.21**	**0.91**	**0.16**	**0.92**	**0.12**	**0.91**	**0.09**
R1	0.93	0.98	0.91	0.71	0.92	0.50	0.93	0.40	0.91	0.30	0.92	0.21
	0.94	**0.40**	**0.93**	**0.30**	**0.9**	**0.23**	**0.9**	**0.18**	**0.94**	**0.14**	**0.92**	**0.10**
R2	0.93	0.98	0.91	0.71	0.92	0.50	0.93	0.40	0.91	0.30	0.91	0.21
	0.94	**0.40**	**0.93**	**0.30**	**0.9**	**0.23**	**0.9**	**0.18**	**0.94**	**0.14**	**0.92**	**0.10**
H	0.93	0.98	0.91	0.71	0.92	0.50	0.93	0.40	0.91	0.30	0.91	0.21
	0.93	**0.40**	**0.92**	**0.29**	**0.9**	**0.233**	**0.9**	**0.18**	**0.93**	**0.13**	**0.93**	**0.10**

resampling into the n (original) sample data, offers results noticeable close to the *Holm* method that is based on the actual construction of the pseudo-population into which to resample.

Among the three proposed probability approximations R-1, R-2 and H, the latter shows more erratic and rather weaker performances than both R-1 and R-2 which, in their turn, appear mostly equivalent. In addition, it is worth notice that our proposed resampling method tends to provide good CIs for quantiles, which are population quantities usually tricky to estimate and yet relevant in practice, for instance in studies on household income distribution and social inequalities.

8. Concluding remarks

In the present paper, a resampling technique for πps sampling designs is presented. Following the classification in Mashreghi et al. (2016), it represents a unified approach to resampling from finite population. On the theoretical ground, due to its relationships with pseudo-population based resampling (cfr. Section 4), it is "asymptotically correct" according to Conti et al. (2020). However, it does not require an explicit construction of a pseudo-population, because bootstrap samples are directly drawn from the original sample on the basis of an appropriate weighting system, so that it is computationally efficient. Real applications of finite population resampling usually involve a form of rounding or re-scaling, either deterministic or based on randomization, that would affect the bootstrap performance and ultimately the expected properties of the released bootstrap estimates. The resampling proposed in this paper does not need any rounding, because it admits an underlying pseudo-population possibly of non-integer size, along with any real value for the bootstrap weights. As a consequence, efficiency gains are expected. Finally, our resampling is very simple to be implemented, since it requires, as a resampling design, a unique basic *ppswor*-type design that is easily implemented in practice.

To be implemented, the resampling scheme here introduced requires the choice of two quantities, namely (i) the number N_i^*, not necessarily integer, of replicates of each sample unit i, and (ii) the drawing probabilities p_i^*. As far as the choice of N_i^*s is concerned, the most natural choice appears to be $N_i^* = \pi_i^{-1}$, that can be implemented even if N_i^*s are not integer. When attention is paid to drawing probabilities, approximations $R-1$ and $R-2$ offer good results, although $R-2$ is slightly heavier from a computational point of view.

The simulation results of Section 7 also add numerical evidence to the theoretical justifications of Section 6, and explain why our methodology outperform Quatember (2014) original proposal as its main competitor.

Appendix A. Proofs

Proof of Proposition 1. Define, as $k \geq 1$,

$$m_{kX}^* = \frac{1}{N^*} \sum_{i \in \boldsymbol{s}} N_i^* x_i^k, \tag{24}$$

(note that $m_{1X}^* = \overline{X}^*$).

Next, let us examine first the ratio $p_i^* / \overline{p}^*$. From (17), (22) and (24), it follows that

$$\begin{aligned}
\frac{p_{(i)}^*}{\overline{p}^*} &= \frac{\frac{1}{N^*} \log\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right)}{\frac{1}{N^*} \sum_{i \in \boldsymbol{s}} N_i^* \left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right) \log\left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right)} \\
&= \frac{\frac{1}{N^*} \left(-f_N^* \frac{x_i}{\overline{X}^*} - \frac{f_N^{*2}}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*3}\right)\right)}{\frac{1}{N^*} \sum_{i \in \boldsymbol{s}} N_i^* \left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right) \left(-f_N^* \frac{x_i}{\overline{X}^*} - \frac{f_N^{*2}}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*3}\right)\right)} \\
&= \frac{\frac{1}{N^*} \left(\frac{x_i}{\overline{X}^*} + \frac{f_N^*}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right)}{\frac{1}{N^*} \sum_{i \in \boldsymbol{s}} N_i^* \left(1 - f_N^* \frac{x_i}{\overline{X}^*}\right) \left(\frac{x_i}{\overline{X}^*} + \frac{f_N^*}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right)} \\
&= \frac{\frac{1}{N^*} \left(\frac{x_i}{\overline{X}^*} + \frac{f_N^*}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right)}{\frac{1}{N^*} \sum_{i \in \boldsymbol{s}} N_i^* \left(\frac{x_i}{\overline{X}^*} - \frac{f_N^*}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right)} \\
&= \frac{\frac{1}{N^*} \left(\frac{x_i}{\overline{X}^*} + \frac{f_N^*}{2} \frac{x_i^2}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right)}{1 - \frac{f_N^*}{2} \frac{m_{2X}^*}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)} .
\end{aligned} \tag{25}$$

Similarly from (16), (17) and (21) it is seen that

$$\frac{\overline{\pi}^*}{\pi^*_{(i)}} = \frac{\sum_{i \in \mathbf{s}} N^*_i \left(f^*_N \frac{x_i}{\overline{X}^*} - f^{*2}_N \frac{x_i^2}{\overline{X}^{*2}} \right)}{f^*_N \frac{x_i}{\overline{X}^*}}$$
$$= \frac{N^*}{x_i} \left(\overline{X}^* - f^*_N \frac{m^*_{2X}}{\overline{X}^*} \right). \tag{26}$$

Now, as a consequence of (25) and (26), it follows

$$\frac{p^*_{(i)}}{\overline{p}^*} \frac{\overline{\pi}^*}{\pi^*_{(i)}} = \frac{1 - f^*_N \frac{m^*_{2X}}{\overline{X}^{*2}} + \frac{f^*_N}{2} \frac{x_i}{\overline{X}^*} + O\left(f^{*2}_N\right)}{1 - \frac{f^*_N}{2} \frac{m^*_{2X}}{\overline{X}^{*2}} + O\left(f^{*2}_N\right)}$$
$$= \frac{1 + f^*_N \left(\frac{x_i}{2\overline{X}^*} - \frac{m^*_{2X}}{\overline{X}^{*2}} \right) + O\left(f^{*2}_N\right)}{1 - \frac{f^*_N}{2} \frac{m^*_{2X}}{\overline{X}^{*2}} + O\left(f^{*2}_N\right)}$$
$$= \left\{ 1 + f^*_N \left(\frac{x_i}{2\overline{X}^*} - \frac{m^*_{2X}}{\overline{X}^{*2}} \right) + O\left(f^{*2}_N\right) \right\} \left(1 + \frac{f^*_N}{2} \frac{m^*_{2X}}{\overline{X}^{*2}} + O\left(f^{*2}_N\right) \right)$$
$$= 1 + \frac{f^*_N}{2} \left(\frac{x_i}{\overline{X}^*} - \frac{m^*_{2X}}{\overline{X}^{*2}} \right) + O\left(f^{*2}_N\right). \tag{27}$$

Finally, from (20), (27) it is not difficult to conclude that

$$\frac{\Delta^{*R}_{ij} - \Delta^{*S}_{ij}}{\Delta^{*R}_{ij}} \sim \left(1 - \frac{\overline{\pi}^* p^*_{(i)}}{\overline{p}^* \pi^{*R}_{(i)}} \right) \left(1 - \frac{\overline{\pi}^* p^*_{(j)}}{\overline{p}^* \pi^{*R}_{(j)}} \right)$$
$$= \frac{f^{*2}_N}{4} \left(\frac{x_i}{\overline{X}^*} - \frac{m^*_{2X}}{\overline{X}^{*2}} \right) \left(\frac{x_j}{\overline{X}^*} - \frac{m^*_{2X}}{\overline{X}^{*2}} \right) + O\left(f^{*3}_N\right),$$

from which (23) follows. □

Appendix B. Additional simulation results

Let $\hat{v}^*$ be the (point) bootstrap variance estimate of a given estimator among the four simulated (see Section 7), and let V be the actual estimator's variance. Thus, the simulation metrics RB and $RRMSE$ are computed as $RB = E_{MC}\left[\hat{v}^* - V\right]/V$ and $RRMSE = \sqrt{E_{MC}\left[\hat{v}^* - V\right]^2}/V$ where E_{MC} indicates average over the 1,000 simulated samples (Monte Carlo runs) for each scenario in Table 1.

Table 6
Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) of the point bootstrap estimate of the standard error of all estimators simulated, for increasing population sizes and two sample fractions 4% **20%**.

N	200		400		800		1200		2400		5000	
	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\hat{\hat{Y}}_{HT}$												
HTB	0.18	1.11	-0.26	0.96	0.22	1.1	0.21	1.11	0.13	0.71	-0.07	0.79
	0.15	**0.93**	**-0.19**	**0.87**	**0.19**	**0.93**	**-0.16**	**0.89**	**-0.1**	**0.66**	**-0.06**	**0.68**
Holm	0.09	1.04	-0.13	1.09	-0.09	1.08	-0.1	0.98	-0.09	0.79	-0.05	0.57
	-0.07	**0.89**	**-0.11**	**0.84**	**-0.08**	**0.93**	**-0.09**	**0.89**	**-0.06**	**0.66**	**-0.05**	**0.45**
Q	0.22	1.03	-0.24	1.2	0.31	1.36	-0.96	1.15	-0.93	1.37	-1.26	1.06
	-0.19	**1.01**	**-0.17**	**0.99**	**0.25**	**1.09**	**-0.69**	**0.99**	**0.74**	**0.99**	**-1**	**1**
R-1	0.11	0.99	-0.2	1	0.13	1.06	0.14	0.95	-0.12	0.96	0.06	0.5
	-0.09	**0.89**	**-0.18**	**0.86**	**-0.12**	**0.97**	**0.11**	**0.91**	**0.08**	**0.69**	**0.05**	**0.46**
R-2	0.13	1.05	0.23	0.94	-0.13	1.17	0.11	0.98	0.1	0.76	0.07	0.53
	-0.11	**0.89**	**-0.18**	**0.83**	**-0.11**	**0.97**	**0.09**	**0.91**	**-0.08**	**0.7**	**-0.06**	**0.46**
H	-0.13	1.04	0.2	1.08	-0.18	0.99	0.1	1	-0.06	0.83	-0.06	0.48
	-0.11	**0.9**	**0.15**	**0.87**	**0.18**	**0.97**	**-0.09**	**0.91**	**-0.07**	**0.7**	**-0.06**	**0.46**

(*continued on next page*)

Table 6 (*continued*)

N	200		400		800		1200		2400		5000	
	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
$\hat{\hat{Y}}_H$												
HTB	-0.92	0.86	-0.86	0.73	-0.93	0.9	1.1	0.73	0.92	0.78	0.4	0.62
	-0.82	**0.77**	**0.65**	**0.54**	**0.71**	**0.69**	**0.79**	**0.54**	**-0.77**	**0.7**	**0.37**	**0.53**
Holm	1.15	0.83	0.81	0.61	0.79	0.73	-0.7	0.71	0.26	0.44	0.1	0.43
	0.85	**0.71**	**0.67**	**0.52**	**-0.68**	**0.63**	**-0.59**	**0.57**	**-0.2**	**0.41**	**-0.08**	**0.42**
Q	1.3	1.07	0.93	0.68	0.99	1.11	-0.78	1.01	-0.63	1.13	1.18	0.92
	0.98	**0.82**	**-0.78**	**0.62**	**-0.86**	**0.97**	**0.79**	**0.93**	**-0.61**	**0.99**	**-0.93**	**0.8**
R-1	-1.06	0.8	0.96	0.62	-0.88	0.76	-0.7	0.76	0.2	0.61	-0.11	0.53
	0.81	**0.72**	**-0.67**	**0.54**	**-0.73**	**0.63**	**-0.71**	**0.57**	**-0.18**	**0.5**	**-0.09**	**0.49**
R-2	-1.02	0.83	0.86	0.75	1.02	0.88	0.76	0.75	0.17	0.8	-0.11	0.57
	0.84	**0.75**	**-0.7**	**0.54**	**-0.73**	**0.63**	**-0.71**	**0.57**	**-0.18**	**0.5**	**-0.09**	**0.48**
H	1.08	0.97	-0.81	0.64	0.98	0.94	-0.68	0.73	0.28	0.53	0.12	0.56
	0.93	**0.79**	**-0.69**	**0.54**	**-0.74**	**0.71**	**-0.7**	**0.55**	**-0.22**	**0.49**	**-0.14**	**0.53**
$\hat{Me}_Y$												
HTB	1.14	1.3	-1.33	1.05	-1.21	0.84	-1.36	0.65	0.78	0.66	-0.32	0.75
	0.97	**1.16**	**1.02**	**1.06**	**1.04**	**0.72**	**0.95**	**0.63**	**0.74**	**0.64**	**0.26**	**0.71**
Holm	1.16	1.27	-1.2	1.02	-1.11	0.88	1.05	0.89	-0.14	0.6	-0.06	0.56
	-0.92	**1.11**	**0.92**	**0.95**	**0.89**	**0.7**	**-0.85**	**0.58**	**-0.11**	**0.51**	**-0.06**	**0.49**
Q	1.07	1.43	-1.31	1.06	1.32	1.11	-1.57	1.15	-0.28	0.81	-0.86	0.92
	0.88	**1.24**	**0.98**	**0.96**	**1.11**	**0.94**	**-1.06**	**0.77**	**-0.26**	**0.79**	**-0.81**	**0.79**
R-1	-1.04	1.36	-1.02	1.09	1.09	0.88	-0.92	0.57	-0.13	0.81	0.06	0.64
	-0.86	**1.1**	**0.86**	**1**	**0.85**	**0.69**	**-0.83**	**0.59**	**-0.14**	**0.7**	**-0.05**	**0.53**
R-2	-0.95	1.27	1.05	1.4	0.94	0.77	0.99	0.76	-0.18	0.73	0.06	0.62
	-0.74	**1.07**	**0.86**	**1**	**0.86**	**0.7**	**-0.89**	**0.59**	**-0.13**	**0.7**	**-0.05**	**0.54**
H	-0.85	1.28	1.02	1.1	1	0.83	-0.91	0.72	-0.17	0.83	-0.11	0.62
	0.65	**1.14**	**-0.85**	**1**	**0.86**	**0.71**	**-0.84**	**0.57**	**-0.14**	**0.68**	**-0.1**	**0.54**
$\hat{Q}3_Y$												
HTB	1.4	1.55	-1.15	1.11	1.07	0.8	0.97	0.74	-0.91	0.79	0.45	0.9
	0.84	**1.32**	**0.98**	**0.85**	**-0.92**	**0.69**	**0.8**	**0.62**	**0.75**	**0.66**	**0.33**	**0.77**
Holm	-1.77	1.65	-0.99	0.92	1.01	0.63	-0.99	0.65	-0.2	0.58	-0.08	0.69
	-0.98	**1.27**	**0.87**	**0.75**	**0.85**	**0.53**	**-0.81**	**0.57**	**-0.12**	**0.54**	**-0.07**	**0.49**
Q	-1.39	1.58	1.21	0.9	1.07	0.9	1.55	0.75	-1.43	1.16	-0.77	0.96
	0.96	**1.38**	**-0.98**	**0.75**	**0.97**	**0.82**	**-1.09**	**0.76**	**-1.03**	**1.06**	**-0.66**	**0.9**
R-1	-0.92	1.45	1.1	0.98	1.13	0.94	-0.83	0.58	-0.3	0.7	0.07	0.49
	-0.75	**1.24**	**-0.86**	**0.79**	**-0.85**	**0.71**	**-0.82**	**0.57**	**-0.22**	**0.54**	**-0.06**	**0.46**
R-2	1.46	1.46	-1.06	0.87	1.02	0.83	-0.9	0.56	0.23	0.62	0.07	0.65
	-0.87	**1.27**	**-0.86**	**0.79**	**-0.87**	**0.71**	-0.82	**0.57**	**-0.21**	**0.54**	-0.06	**0.46**
H	-1.01	1.49	-1.19	0.83	-1.19	0.97	-1.04	0.67	-0.24	0. 57	-0.11	0.59
	-0.87	**1.28**	**-0.82**	**0.78**	**-0.88**	**0.72**	**0.87**	**0.56**	**-0.18**	**0.52**	**-0.09**	**0.5**

Appendix C. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107366.

References

Antal, E., Tillé, Y., 2011. A direct bootstrap method for complex sampling designs from a finite population. J. Am. Stat. Assoc. 106 (494), 534–543.

Bondesson, L., Traat, I., Lundqvist, A., 2006. Pareto sampling versus Sampford and conditional Poisson sampling. Scand. J. Stat. 33, 699–720.

Chen, S., Haziza, D., Lèger, C., Mashregi, Z., 2019. Pseudo-population bootstrap methods for imputed survey data. Biometrika 106, 369–384.

Conti, P.L., Marella, D., Mecatti, F., Andreis, F., 2020. A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Bernoulli 26, 1044–1069.

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

Hájek, J., 1973. Asymptotic theories of sampling with varying probabilities without replacement. In: Hájek, J. (Ed.), Proceedings of the Prague Symposium on Asymptotic Statistics, vol. 2. Charles University, Prague, pp. 127–138.

Hájek, J., 1981. Sampling from a Finite Population. Marcel Dekker, New York.

Holmberg, A., 1998. A bootstrap approach to probability proportional-to-size sampling. In: Proceedings of the ASA Section on Survey Research Methods, pp. 378–383.

Kelley, C.T., 1995. Iterative Methods for Linear and Nonlinear Equations. SIAM, Philadelphia.

Mashreghi, Z., Haziza, D., Léger, C., 2016. A survey of bootstrap methods in finite population sampling. Stat. Surv. 10, 1–52.

Quatember, A., 2014. The finite population bootstrap - from the maximum likelihood to the Horvitz-Thompson approach. Aust. J. Stat. 43, 93–102.

Quatember, A., 2015. Pseudo-Populations - A Basic Concept in Statistical Surveys. Springer Verlag, New York.

Ranalli, M.G., Mecatti, F., 2012. Comparing recent approaches for bootstrapping sample survey data: a first step towards a unified approach. In: Proceedings of the ASA Section on Survey Research Methods, pp. 4088–4099.

Rao, J.N.K., Wu, C.F.J., 1988. Resampling inference with complex survey data. J. Am. Stat. Assoc. 83, 231–241.

Rosén, B., 1972. Asymptotic theory for successive sampling with varying probabilities without replacement I, II. Ann. Math. Stat. 43, 373–397. 748–776.
Rosén, B., 1997. On sampling with probability proportional to size. J. Stat. Plan. Inference 62, 159–191.
Rosén, B., 2000. On inclusion probabilities for order πps sampling. J. Stat. Plan. Inference 90, 117–143.