Bernoulli **26**(2), 2020, 1044–1069
https://doi.org/10.3150/19-BEJ1138

A unified principled framework for resampling based on pseudo-populations: Asymptotic theory

PIER LUIGI CONTI[1], DANIELA MARELLA[2], FULVIA MECATTI[3] and FEDERICO ANDREIS[4]

[1]*Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le A. Moro, 5, 00185 Roma, Italy. E-mail: pierluigi.conti@uniroma1.it*
[2]*Dipartimento di Scienze della Formazione, Università Roma Tre, Via D. Manin, 53, 00185 Roma, Italy. E-mail: daniela.marella@uniroma3.it*
[3]*Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, Via Bicocca degli Arcimboldi, 8, 20126 Milano, Italy. E-mail: fulvia.mecatti@unimib.it*
[4]*Faculty of Health Sciences and Sport, University of Stirling, Pathfoot Building, Stirling FK9 4LA, Scotland, UK. E-mail: federico.andreis@stir.ac.uk*

In this paper, a class of resampling techniques for finite populations under πps sampling design is introduced. The basic idea on which they rest is a two-step procedure consisting in: (i) constructing a "pseudo-population" on the basis of sample data; (ii) drawing a sample from the predicted population according to an appropriate resampling design. From a logical point of view, this approach is essentially based on the *plug-in* principle by Efron, at the "sampling design level". Theoretical justifications based on large sample theory are provided. New approaches to construct pseudo populations based on various forms of calibrations are proposed. Finally, a simulation study is performed.

Keywords: πps sampling designs; bootstrap; calibration; confidence intervals; finite populations; resampling; variance estimation

1. Introduction

The use of resampling methods in survey sampling has a long history, and several different techniques have been proposed in the literature. The common starting point consists in observing that the classical bootstrap method, as proposed by Efron [22], does not work in survey sampling, because of the dependence among units due to the sampling design itself.

Adaptations taking into account the non *i.i.d.* nature of the data are required when the sample is collected through a general sampling design, possibly assigning different probability to every population unit to be included in the sample. The literature on resampling from finite populations is mainly devoted to estimate variances of estimators; cfr. Mashreghi *et al.* [34]. The main approaches are essentially two: *ad hoc* approaches and *plug in* approaches (cfr. Ranalli and Mecatti [37], Chauvet [15] and references therein).

The basic idea of *ad hoc* approaches consists in resampling from the original sample through a special design, that accounts for the dependence among units. This approach is pursued in McCarthy and Snowden [35], Rao and Wu [38], where the re-sampled data produced by the "usual"

i.i.d. bootstrap are properly rescaled, as well as in Sitter [39], Beaumont and Patak [2], Chatterjee [14], Conti and Marella [18], where a "rescaled bootstrap process" based on asymptotic results is proposed. Among the *ad hoc* approaches we also quote the recent paper by Antal and Tillé [1], where an ingenious mixed resampling design is proposed to account for the dependence among observations.

Plug-in approaches are based on the idea of "expanding" the sample to a "pseudo-population" that plays the role of a "surrogate" (actually a prediction) of the original one. Then, bootstrap samples are drawn from such a pseudo-population according to some appropriate resampling design. The most intuitive choice consists in using the *same* sampling design used to draw the original sample from the population; cfr. Gross [25], Chao and Lo [13], Booth *et al.* [10], Holmberg [28], Chauvet [15], as well as Mashreghi *et al.* [34]. Applications of resampling methods based on pseudo-populations are in Marella and Vicard [33], where the structural learning of graphical models with data coming from a complex sample survey is dealt with, in Conti *et al.* [20], where a statistical matching problem is studied, and in Lahiri [30].

Virtually all resampling techniques proposed for finite populations rest on the same justification: in case of linear statistics, the variance of the resampled statistic should match (or should be very close to) the "usual" variance estimator, possibly with approximated forms of the second order inclusion probabilities; cfr. Antal and Tillé [1]. This is far from the arguments commonly used to justify the classical bootstrap and its variants, that are based on asymptotic considerations involving the whole sampling distribution of a statistic (cfr., for instance, Bickel and Freedman [7] and Lahiri [31]): the asymptotic distribution of a bootstrapped statistic should coincide with that of the "original" statistic. This argument is actually used in Conti and Marella [18].

In the present paper, a class of resampling techniques for finite populations is proposed. It is based on a two-phase procedure. In the first phase, a pseudo-population, that can be viewed as a prediction of the population, is constructed. In the second phase, a (re)sample is drawn from the pseudo-population. In a broad sense, this approach parallels the *plug-in principle* by Efron [23]. The pseudo-population is plugged in the sampling process, and acts as a surrogate of the actual finite population. In other terms, the predicted population mimics the real population, and the (re)sampling process from the predicted population mimics the (original) sampling process from the real population. From a formal point of view, the main justification of the whole procedure is based on large sample arguments. In this sense, the approach pursued in the present paper offers a principled framework for resampling from finite populations that parallels the arguments used for classical Efron's bootstrap of *i.i.d.* data. For this reason, some preliminary developments of large sample theory for finite populations are needed. The asymptotic framework considered here is essentially that in Isaki and Fuller [29], where a sequence of finite populations of increasing size is considered, and the sample size correspondingly increases. Furthermore, high entropy sampling designs are considered, similar to those studied in Conti [17], Conti and Marella [18], but with an important addition: the possible relationships between the variable of interest and the design variables are explicitly taken into account. This dramatically changes the asymptotic results in Conti [17]. As a matter of fact, the resampling method defined in Conti and Marella [18], based on rescaling Efron's bootstrap, does not apply when there is dependence between the variable of interest and the design variables, which is often the case in real applications.

The results in Propositions 1–3 have strong connections with other results recently appeared in the literature: cfr. Boistard *et al.* [8], Bertail *et al.* [6]. In the first paper, the authors study the

asymptotic behaviour of the process (25), under fairly general conditions on the sampling design, including high entropy sampling designs as special cases. In Bertail *et al.* [6], empirical processes indexed by classes of functions are studied for high-entropy sampling designs. However, there are important differences, that should be highlighted. In the present paper, we study our functional central limit theorems conditionally on the realization of the superpopulation, as both the sample and population sizes go to infinity. In a sense, they are in the same spirit as limit results in Bickel and Freedman [7]. Functional central limit theorems in Boistard *et al.* [8] are studied unconditionally. The need to establish conditional results is due to two reasons. First of all, here we consider a design-based approach, where the only source of variability is the sampling design: see, *for example*, Cassel *et al.* [12], Ch. 1. In the second place, the asymptotic study of resampling procedures requires conditioning. The results in Bertail *et al.* [6] are conditional results, in the same spirit as the present paper. However, they are established for Horvitz–Thompson estimator of the population distribution function, and not for the Hájek estimator as in the present case. Furthermore, regularity assumptions in Bertail *et al.* [6] are slightly different from ours.

The paper is organized as follows. In Section 2, basic notions are introduced, and in Section 3 different strategies to construct pseudo-populations are discussed. Section 4 contains the main technical assumptions on which the paper rests. Sections 5, 6 are devoted to asymptotic results for a wide class of estimators of appropriate population parameters. Section 7 describes the proposed resampling and the basic theoretical results. Properties of calibrated pseudo-populations are studied in Section 8. In Section 9, different methods to construct pseudo-populations are compared *via* a Monte Carlo simulation study. Conclusions are provided in Section 10. Technical lemmas and proofs are gathered in Appendix [19].

2. Basics aspects and notation

Let $\mathcal{U}_N$ be a finite population of size N. The character of interest is denoted by $\mathcal{Y}$, and its value for unit i by y_i; furthermore, let $\boldsymbol{y}_N = (y_1, \ldots, y_N)$.

A sample $\boldsymbol{s}$ is a subset of $\mathcal{U}_N$. Denote by D_i the sample membership indicator of unit i, namely a Bernoulli random variable (r.v.), such that i is (is not) in the sample $\boldsymbol{s}$ whenever $D_i = 1$ ($D_i = 0$); clearly, $\boldsymbol{s} = \{i \in \mathcal{U}_N : D_i = 1\}$. Denote further by $\boldsymbol{D}_N$ the N-dimensional r.v. of components $(D_1, \ldots, D_N)$. 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. a sampling design P.

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 suffix P denotes the sampling design used to select the sample $\boldsymbol{s}$. The sample size is $n_s = D_1 + \cdots + D_N$.

The first order inclusion probabilities are frequently taken 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 ($i = 1, \ldots, N$). 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. The symbol $\boldsymbol{x}_N$, from now on, will denote the sequence $(x_1, \ldots, x_N)$.

For each unit i, let p_i be a positive number, with $p_1 + \cdots + p_N = n$. The *Poisson sampling design* (Po, for short) with parameters $p_1, \ldots, p_N$ is characterized by the independence of the r.v.s D_is, with $\Pr_{\text{Po}}(D_i = 1) = p_i$. In symbols

$$\Pr_{\text{Po}}(\boldsymbol{D}_N) = \prod_{i=1}^{N} p_i^{D_i} (1 - p_i)^{1 - D_i}.$$

The *rejective sampling*, or *normalized conditional Poisson sampling* (cfr. Hájek [26], Tillé [41]) is obtained from the Poisson sampling by conditioning w.r.t. $n_s = n$. Using the suffix P_R to denote the rejective sampling design, $E_{P_R}[D_i]$ is *not* generally equal to p_i, although they are asymptotically equivalent, as N and n increase (Hájek [26]). In Chen *et al.* [16] an algorithm is proposed to compute p_is in terms of π_is for the conditional Poisson sampling.

The rejective sampling design is characterized by a fundamental property: it possesses maximum entropy among all sampling designs of fixed size and fixed first order inclusion probabilities (as shown in Hájek [27]), where the entropy of a sampling design P is

$$H(P) = -E_P\big[\log \Pr_P(\boldsymbol{D}_N)\big] = -\sum_{D_1, \ldots, D_N} \Pr_P(\boldsymbol{D}_N) \log\big(\Pr_P(\boldsymbol{D}_N)\big).$$

The *Hellinger distance* between a sampling design P and the rejective design is defined as

$$d_H(P, P_R) = \sum_{D_1, \ldots, D_N} \Big(\sqrt{\Pr_P(\boldsymbol{D}_N)} - \sqrt{\Pr_{P_R}(\boldsymbol{D}_N)}\Big)^2. \tag{1}$$

3. Pseudo-population: Construction based on calibration

The class of resampling techniques we consider rests on a two-phase procedure. In the first phase, on the basis of the sampling data a pseudo-population, *that is*, a design-based predictor of the actual population, is constructed. In the second phase, a sample of size n (the same as the "original" one) is drawn from the pseudo-population, according to a πps sample design P^* (the resampling design) with inclusion probabilities again proportional to x_is. Intuition suggests to use a resampling design of the same type as the sampling design used to draw the sample $\boldsymbol{s}$ from the population. This point will be discussed later, in Section 7.2.

Formally speaking, a pseudo-population $\mathcal{U}^*_{N^*}$ is

$$\big\{\big(N_i^* D_i, y_i, x_i\big); i = 1, \ldots, N\big\} \tag{2}$$

where N_i^*s are integer-valued r.v.s, with (joint) probability distribution P_{pred}. In practice, (2) means that $N_i^* D_i$ population units are predicted to have y-value equal to y_i and x-value equal to x_i, for each sample unit i. In the sequel, the symbols y_k^*, x_k^* will be used to denote the y-value and x-value of unit k of the pseudo-population, respectively. The quantity

$$N^* = \sum_{i=1}^{N} N_i^* D_i. \tag{3}$$

is the size of the pseudo-population.

A relevant aspect that would potentially affect the performance of resampling, is how the pseudo-population is constructed. The intuition behind pseudo-populations is simple: the pseudo-population should be as "similar" as possible to the actual finite population. In a sense, the pseudo-population should be somehow calibrated w.r.t. the population. Such an intuition can be put into practice in several ways. In the present section, some classical proposals are reviewed, and some new proposals based on different calibration approaches are illustrated.

3.1. Holmberg pseudo-population

Following the popular Horvitz–Thompson (HT) approach to πps sampling and estimation, each unit $i \in \boldsymbol{s}$ should be "predicted" in $\mathcal{U}^*_{N^*}$ a number of times equal to its design weight π_i^{-1}, provided they are all integers. For the general non-integer case the following strategy has been proposed by Holmberg [28]. Let $r_i = \pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$, and consider independent Bernoulli r.v.s ϵ_is with $\Pr(\epsilon_i = 1 | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N) = r_i$. A HT pseudo-population is constructed by replicating every sampled unit $i \in \boldsymbol{s}$ $N_i^{*\mathrm{HT}} = \lfloor \pi_i^{-1} \rfloor + \epsilon_i$ times, with corresponding values y_i, x_i. The size of a HT pseudo-population is $N^{*\mathrm{HT}} = \sum_{i=1}^{N} N_i^{*\mathrm{HT}} D_i$, which is not generally equal to N.

3.2. Multinomial pseudo-population

For $k = 1, \ldots, N$, perform independent trials consisting in choosing a unit from the original sample, where each unit i is selected with probability

$$\pi_i^{-1} / \sum_{j \in \boldsymbol{s}} \pi_j^{-1} = x_i^{-1} / \sum_{j \in \boldsymbol{s}} x_j^{-1}.$$

If at trial k unit i is selected, unit k of the pseudo-population will take values $y_k^* = y_i$ and $x_k^* = x_i$. If $N_i^{*\mathrm{MUL}}$, $i \in \boldsymbol{s}$, is the number of replications of unit i in the pseudo-population, then (conditionally on $\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N$) the r.v. $(N_i^{*\mathrm{MUL}}; i \in \boldsymbol{s})$ possesses a multinomial distribution, with

$$E\left[N_i^{*\mathrm{MUL}} | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N\right] = N D_i \pi_i^{-1} / \sum_{j=1}^{N} D_j \pi_j^{-1} \tag{4}$$

$$V\left(N_i^{*\mathrm{MUL}} | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N\right) = N \left(D_i \pi_i^{-1} \Big/ \sum_{j=1}^{N} D_j \pi_j^{-1} \right) \times \left(1 - D_i \pi_i^{-1} \Big/ \sum_{j=1}^{N} D_j \pi_j^{-1} \right) \tag{5}$$

$$C\left(N_i^{*\mathrm{MUL}}, N_h^{*\mathrm{MUL}} | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N\right) = -N D_i D_h \pi_i^{-1} \pi_h^{-1} \Big/ \left(\sum_{j=1}^{N} D_j \pi_j^{-1} \right)^2, \quad h \neq i. \tag{6}$$

This approach goes essentially back to Sverchkov and Pfeffermann [40] and guarantees by construction a pseudo-population calibrated w.r.t. the population size.

3.3. Conditional Poisson pseudo-population

The Holmberg scheme in Section 3.1 is essentially based on drawing a Poisson sample from $\boldsymbol{s}$, with inclusion probabilities r_is, $i \in \boldsymbol{s}$. A simple idea to calibrate such scheme in order to produce a pseudo-population of exactly N units, consists in defining the quantities

$$\tau_i = N \frac{\pi_i^{-1}}{\sum_{k \in \boldsymbol{s}} \pi_k^{-1}} - \left\lfloor N \frac{\pi_i^{-1}}{\sum_{k \in \boldsymbol{s}} \pi_k^{-1}} \right\rfloor, \quad i \in \boldsymbol{s}$$

and in drawing from $\boldsymbol{s}$ a sample $\boldsymbol{s}_0$ of

$$n_0 = \sum_{i \in \boldsymbol{s}} \tau_i = N - \sum_{i \in \boldsymbol{s}} \left\lfloor N \frac{\pi_i^{-1}}{\sum_{k \in \boldsymbol{s}} \pi_k^{-1}} \right\rfloor$$

units, according to a conditional Poisson sampling design with first order inclusion probabilities τ_is.

For each unit $i \in \boldsymbol{s}$, let ϵ_i be equal to 1 iff i is in $\boldsymbol{s}_0$, and $\epsilon_i = 0$ otherwise. Each unit i of the original sample is replicated in the pseudo-population exactly

$$N_i^{*\mathrm{CPP}} = \left\lfloor N \frac{\pi_i^{-1}}{\sum_{k \in \boldsymbol{s}} \pi_k^{-1}} \right\rfloor + \epsilon_i \tag{7}$$

times.

3.4. Double-calibrated pseudo-population

The conditional Poisson pseudo-population illustrated in Section 3.3 is calibrated w.r.t. the population size N, but not w.r.t. the mean of the auxiliary variable X. A natural idea would consist in modifying $N_i^{*\mathrm{CPP}}$ defined by equation (7) in order to satisfy a further constraint: *the mean of X in the pseudo-population is equal to the mean of X in the actual population*.

Take $N_i^{*\mathrm{CPP}}$, $i \in \boldsymbol{s}$, as an "initial" solution for replicates of sample units in the pseudo-population, and let further N^* be the total number of pseudo-population units (cfr. (3)), and

$$\overline{X}_N = N^{-1} \sum_{i=1}^{N} x_i, \qquad \overline{X}^*_{N^*} = N^{*-1} \sum_{i=1}^{N} N_i^* x_i D_i. \tag{8}$$

The basic idea is to choose pseudo-population replicates that satisfy both constraints on population size and mean of X, and that are as close as possible to the initial $N_i^{*\mathrm{CPP}}$s. More formally, the pseudo-population replicates are taken equal to $N_i^{*\mathrm{DCal}}$s, the solution of the following

quadratic problem:

$$\begin{cases} \min \sum_{i \in \boldsymbol{s}} \left(N_i^* - N_i^{*\mathrm{CPP}}\right)^2 \\ N^* = N \\ \overline{X}_{N^*}^* = \overline{X}_N \\ N_i^* \geq 1 \end{cases} \tag{9}$$

The values $N_i^{*\mathrm{DCal}}$s obtained by solving (9) are not necessarily integer-valued. In order to obtain integer values, it is enough to apply to $N_i^{*\mathrm{DCal}}$s a randomization device similar to that of CPP pseudo-population described in Section 3.3.

3.5. Hot-deck pseudo-population

The basic idea of the calibrated pseudo-population introduced in Section 3.4 consists in constructing a pseudo-population that is "similar" for some characteristics of the auxiliary variable X w.r.t. the original population. This idea is pursued by taking only the sample x_is values. When x_is are available for all population units, the notion of pseudo-population can be extended by considering predictors of the form $\{(x_i^*, y_i^*), i = 1, \ldots, N\}$, where $x_i^* = x_i$ for every unit $i = 1, \ldots, N$ and $y_i^* = \widehat{y}_i =$ imputed value for y_i, according to hot-deck imputation. In detail, the hot-deck pseudo-population is composed by N units, i.e. $\mathcal{U}_{N^*}^* = \mathcal{U}_N^*$. A pair of values (x_i^*, y_i^*) corresponds to each unit $i \in \mathcal{U}_N^*$, with

$$x_i^* = x_i, \quad i = 1, \ldots, N \tag{10}$$

$$y_i^* = \begin{cases} y_i & \text{if } i \in \boldsymbol{s} \\ y_j & \text{with } j = \underset{j \in \boldsymbol{s}}{\operatorname{argmin}} |x_j - x_i| \text{ if } i \in \mathcal{U}_N^* \setminus \boldsymbol{s}. \end{cases} \tag{11}$$

In other terms, for each unit $i \in \mathcal{U}_N^*$ a *donor unit* $j(i)$ is chosen, such that

$$j(i) := \begin{cases} i & \text{if } i \in \boldsymbol{s} \\ |x_{j(i)} - x_i| = \min\limits_{j \in U_N \setminus \boldsymbol{s}} |x_j - x_i| & \text{if } i \in \mathcal{U}_N^* \setminus \boldsymbol{s}. \end{cases}$$

The value y_i^* for unit i is then taken equal to those of its donor, leading to a pseudo-population which is calibrated by construction w.r.t. both population size N and the entire distribution of the auxiliary variable X.

4. Basic assumptions

Denote by $\mathcal{Y}$ the character of interest, and let y_i be its value for unit i. $\mathcal{T}_1, \ldots, \mathcal{T}_L$ are the design variables, and $t_{i1}, \ldots, t_{iL}$ are their values for unit i. The design variables may include strata indicators, as well as variables measuring cluster and unit characteristics (cfr. Pfeffermann [36]).

They are used to construct the sampling design, and to compute the sampling weights, *that is*, the reciprocals of the first order inclusion probabilities.

The basic assumptions on which the present paper relies are listed below.

A1. $(\mathcal{U}_N; N \geq 1)$ is a sequence of finite populations of increasing size N.

A2. For each N, $(y_i, t_{i1}, \ldots, t_{iL})$, $i = 1, \ldots, N$ are realizations of a superpopulation model $\{(Y_i, T_{i1}, \ldots, T_{iL}), i = 1, \ldots, N\}$ composed by *i.i.d.* $(L+1)$-dimensional r.v.s. The symbol $\mathbb{P}$ denotes the (superpopulation) probability distribution of r.v.s $(Y_i, T_{i1}, \ldots, T_{iL})$s, and $\mathbb{E}$, $\mathbb{V}$ are the corresponding mean and variance, respectively.

A3. For each population $\mathcal{U}_N$, sample units are selected according to a sample design with positive first order inclusion probabilities $\pi_1, \ldots, \pi_N$, and fixed sample size $n = \pi_1 + \cdots + \pi_N$. The first order inclusion probabilities are taken proportional to $x_i = h(t_{i1}, \ldots, t_{iL})$, $h(\cdot)$ being an arbitrary, strictly positive function. To avoid complications in the notation, we will assume that $\pi_i = n x_i / \sum_{i=1}^{N} x_i$ for each unit i.

Although the sample size n, the inclusion probabilities π_is, and the r.v.s D_is, as well, depend on N, in order to use a simple notation the symbols n, π_i, D_i are used, instead of the more complete n_N, $\pi_{i,N}$, $D_{i,N}$. It is also assumed that

$$\lim_{N,n\to\infty} \mathbb{E}\big[\pi_i(1-\pi_i)\big] = d > 0. \tag{12}$$

A4. The sample size n increases as the population size N does, with

$$\lim_{N\to\infty} \frac{n}{N} = f, \quad 0 < f < 1. \tag{13}$$

A5. For each population $(\mathcal{U}_N; N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities $\pi_1, \ldots, \pi_N$, and let P be the actual sampling design (with the same inclusion probabilities). Then

$$d_H(P, P_R) \to 0 \quad \text{as } N \to \infty, \text{ a.s.-}\mathbb{P}.$$

A6. $\mathbb{E}[X_1^2] < \infty$, so that the quantity in (12) is equal to:

$$d = f\left(1 - \frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2}\right) + f(1-f)\frac{\mathbb{E}[X_1^2]}{\mathbb{E}[X_1]^2} > 0. \tag{14}$$

Cfr. Further Lemma 1 in the Appendix.

Assumptions A2, A3 allow one to take into account the possible dependence between the design variables and the study variable. Of course, this is a key motivation for using non-simple, probability-proportional-to-size designs (dubbed πps sampling designs), where the dependence between X_is and Y_is is important for the efficiency of the estimation of the population mean (and other population parameters, as well). Notice that assumptions A2, A3 do not limit the kind of dependence between X_is and Y_is, that can be completely general.

Assumption A2 is not as restrictive as it could appear at a first glance. For instance, it allows for stratification in a simple way. Suppose there is a unique design variable T, taking L values $t_{(1)}, \ldots, t_{(L)}$. The stratum l is composed by all units for which T takes the value $t_{(l)}$

($l = 1, \ldots, L$), and the distribution of Y_i in each stratum, *i.e.* the distribution of Y_i conditionally on $T_i = t_{(l)}$, may vary across strata.

Remark. Assumption A2 is not necessary for the validity of theoretical results in Sections 5–7, and it is stated in the present form only for the sake of simplicity. In fact, assumption A2 is used to prove Lemmas 1–4, that involve the use of the (strong) law of large numbers for appropriate functions of (Y_i, X_i). But independence assumption is not necessary for the validity of the strong law of large numbers. Suppose, for instance, that the population units are clustered into M clusters, where cluster m is of size N_m, $m = 1, \ldots, M$. If the r.v.s (Y_i, X_i)s are correlated within clusters and independent across clusters, and if, as $N \to \infty$, $M \to \infty$ and N_ms remain bounded, then Lemmas 1–5 still hold, as well as all other results of the paper.

An obvious example of sampling designs satisfying A3 are πps sampling designs, where the first order inclusion probability of unit i is proportional to the value of a size measure. Another elementary example is the stratified design. Assume that the population is subdivided into L strata, composed by $N_1, \ldots, N_L$ units, respectively ($N_1 + \cdots + N_L = N$). Let further $w_l = N_l/N$, and let $g_1, \ldots, g_L$ be arbitrary positive numbers such that $g_1 + \cdots + g_L = 1$. The stratified design drawing (by simple random sampling) $n_l = ng_l$ units from stratum l ($= 1, \ldots, L$) can be considered as a special πps sampling design where the first order inclusion probability for unit i is taken proportional to an auxiliary variable (acting as a size measure) x_i defined as

$$x_i = \frac{g_l}{w_l} \quad \text{if unit } i \text{ is within stratum } l. \tag{15}$$

In fact, from (15) it easily follows that

$$\pi_i = \frac{ng_l}{Nw_l} = \frac{n_l}{N_l} \quad \text{if unit } i \text{ is within stratum } l. \tag{16}$$

In particular, if $g_l = w_l$, then the sampling design reduces to stratified proportional sampling.

As discussed in Conti [17], assumption A5 implies that the Kullback–Leibler divergence of the actual sampling design P w.r.t. the rejective design

$$\Delta_{\mathrm{KL}}(P \| P_R) = H(P_R) - H(P) \tag{17}$$

tends to zero as both n, N increase. Hence, the sampling designs satisfying assumption A5 are essentially "high entropy", single-stage, sampling designs. The importance of the high entropy property of sampling designs is discussed in Brewer and Donadio [11], Grafström [24] and references therein. Examples of sampling designs satisfying A5, as shown in Berger [3], Berger [4], Berger [5], are simple random sampling, Rao–Sampford design, Chao design, stratified design (with bounded number of strata). The systematic sampling design does not satisfy A5, due to its low entropy. However, the randomised systematic sampling design is a high entropy design satisfying A5.

The *population distribution function* (p.d.f., for short) is:

$$F_N(y) = \frac{1}{N}\sum_{i=1}^{N} I_{(y_i \le y)}, \quad y \in \mathbb{R} \tag{18}$$

where the indicator function $I_{(y_i \le y)}$ is equal to 1 if $y_i \le y$, and is equal to 0 otherwise.

A *finite population parameter* is a functional (not necessarily real-valued) of the p.d.f.:

$$\theta_N = \theta(F_N). \tag{19}$$

The simplest approach to estimate a finite population parameter of the form (19) consists in estimating first the p.d.f. (18), and then in replacing F_N in (19) by such an estimate. As an estimator of the p.d.f. (18) we consider here the Hájek estimator:

$$\widehat{F}_H(y) = \frac{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i I_{(y_i \le y)}}{\sum_{i=1}^{N} \frac{1}{\pi_i} D_i} \tag{20}$$

which is a proper distribution function. It can be considered as the "finite population version" of the empirical distribution function, that plays a fundamental role in nonparametric statistics. The finite population parameter (19) is then estimated by

$$\widehat{\theta}_H = \theta(\widehat{F}_H). \tag{21}$$

In a sense, (21) is the "finite population version" of *statistical functionals*.

The main task of Sections 5, 6 is to study the asymptotic properties of (20), (21), respectively. In the sequel, the joint superpopulation d.f. of (Y_i, X_i) will be denoted by

$$H(y, x) = \mathbb{P}(Y_i \le y, X_i \le x) \tag{22}$$

and the marginal superpopulation d.f.s of Y_i and X_i by

$$F(y) = \mathbb{P}(Y_i \le y) = H(y, +\infty), \qquad G(x) = \mathbb{P}(X_i \le x) = H(+\infty, x), \tag{23}$$

respectively. Furthermore, the notation

$$K_\alpha(y) = \mathbb{E}\big[\, X_1^\alpha | Y_1 \le y \big], \quad y \in \mathbb{R}, \alpha = 0, \pm 1, \pm 2 \tag{24}$$

will be used. Note that $K_\alpha(+\infty) = \mathbb{E}[X_1^\alpha]$.

5. Estimating population distribution function

The goal of the present section is to derive the limiting distribution of the Hájek estimator (20), as the sample size and the population size increase. To this purpose, consider the stochastic process $W_N^H = (W_N^H(y);\, y \in \mathbb{R})$, where

$$W_N^H(y) = \sqrt{n}\big(\widehat{F}_H(y) - F_N(y)\big); \quad y \in \mathbb{R}. \tag{25}$$

It can be viewed as the finite population sampling version of the well-known empirical process. The main result of the present section is Proposition 1, that establishes the weak convergence of W_N^H to a Gaussian limiting process. Proposition 1 is in spirit similar to the main result in Conti [17], but with fundamental differences that will be stressed in the sequel.

Before stating Proposition 1, we stress that in our asymptotic approach the actual population y_is and x_is values are considered as *fixed*. The only source of variability is the sampling design, namely $\boldsymbol{D}_N$. If we let the population size N go to infinity, we must also consider corresponding sequences $\boldsymbol{y}_\infty = (y_1, y_2, \dots)$, $\boldsymbol{x}_\infty = (x_1, x_2, \dots)$ of y_is and x_is values. The actual $\boldsymbol{y}_N = (y_1, \dots, y_N)$, $\boldsymbol{x}_N = (x_1, \dots, x_N)$ are the segments of the first N y_is, x_is in the sequences $\boldsymbol{y}_\infty$, $\boldsymbol{x}_\infty$, respectively. As N increases, $\boldsymbol{y}_N$ tends to $\boldsymbol{y}_\infty$ and $\boldsymbol{x}_N$ tends to $\boldsymbol{x}_\infty$. By A2, $\boldsymbol{y}_\infty, \boldsymbol{x}_\infty$ live in a probability space $((\mathbb{R}^2)^\infty, \mathcal{B}(\mathbb{R}^2)^\infty, \mathbb{P}^\infty)$, where $\mathcal{B}(\mathbb{R}^2)^\infty$ is the product Borel σ-field over $(\mathbb{R}^2)^\infty$, and $\mathbb{P}^\infty$ is the product measure on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R})^\infty)$ generated by $\mathbb{P}$. The probability statements we consider are of the form $\Pr_P(\cdot|\boldsymbol{y}_N, \boldsymbol{x}_N)$, with N going to infinity. Conditioning w.r.t. $\boldsymbol{y}_N, \boldsymbol{x}_N$ means that y_is and x_is are considered as fixed (although produced by a superpopulation model). The suffix P means that the probability refers to the sampling design. The results we will obtain hold for "almost all" sequences $\boldsymbol{y}_\infty, \boldsymbol{x}_\infty$ that the superpopulation model in A2 can produce, *that is*, for a set of sequences having $\mathbb{P}^\infty$-probability 1. With a slight lack of precision, but more simply and intuitively, in the sequel we will use the expression "for almost all y_is, x_is values".

Proposition 1. *If the sampling design P satisfies assumptions* A1–A6, *with $\mathbb{P}$-probability* 1, *conditionally on $\boldsymbol{y}_N, \boldsymbol{x}_N$ the sequence $(W_N^H; N \geq 1)$, converges weakly, in $D[-\infty, +\infty]$ equipped with the Skorokhod topology, to a Gaussian process $W^H = (W^H(y); y \in \mathbb{R})$ with zero mean function and covariance kernel*

$$
\begin{aligned}
C^H(y,t) = f&\left\{\frac{\mathbb{E}[X_1]}{f} K_{-1}(y \wedge t) - 1\right\} F(y \wedge t) \\
&- \frac{f^3}{d}\left(1 - \frac{K_1(y)}{\mathbb{E}[X_1]}\right)\left(1 - \frac{K_1(t)}{\mathbb{E}[X_1]}\right) F(y)F(t) \\
&- f\left\{\frac{\mathbb{E}[X_1]}{f}\big(K_{-1}(y) + K_{-1}(t) - \mathbb{E}\big[X_1^{-1}\big] - 1\big)\right\} F(y)F(t), \qquad (26)
\end{aligned}
$$

with d given by (12), *and $a \wedge b = \min(a, b)$.*

The covariance kernel (26) implies expectation w.r.t. the superpopulation model. This does not contradict the consideration of sampling design probabilities conditionally on $\boldsymbol{y}_N, \boldsymbol{x}_N$, but it is only a consequence of the strong law of large numbers. The set of sequences $\boldsymbol{y}_\infty, \boldsymbol{x}_\infty$ for which Proposition 1 holds possesses $\mathbb{P}^\infty$-probability 1, and is determined by the strong law of large numbers.

When X_i and Y_i are independent, the covariance kernel (26) reduces to

$$
f(A-1)\big(F(y \wedge t) - F(y)F(t)\big)
$$

where

$$A = \frac{\mathbb{E}[X_1]}{f}\mathbb{E}\big[X_1^{-1}\big] \tag{27}$$

is, with $\mathbb{P}$-probability 1, the limit of

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}$$

as N goes to infinity. Taking into account that $u \wedge v - uv$ is the covariance kernel of a Brownian bridge $B = (B(t); 0 \le t \le 1)$ (*i.e.*, a Wiener process tied down at 1), we have thus proved the following corollary of Proposition 1.

Corollary 1. *If the sampling design P satisfies assumptions* A1–A6, *and if X_i and Y_i are independent, with $\mathbb{P}$-probability* 1, *conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$ the sequence $(W_N^H; N \ge 1)$, converges weakly, in $D[-\infty, +\infty]$ equipped with the Skorokhod topology, to a Gaussian process that can be represented in the form*

$$\big(f(A-1)B\big(F(y)\big); y \in \mathbb{R}\big) \tag{28}$$

as N goes to infinity, where B is a Brownian bridge and A is given by (27).

Corollary 1 essentially coincides with Proposition 2 in Conti [17]. Proposition 1 is new. Due to the choice of the inclusion probabilities in A3, *that is*, $\pi_i \propto x_i$, and due to the possible dependence between X_i and Y_i (that usually comes true in practice), the limiting Gaussian process is *not* proportional to a Brownian bridge. Proposition 1 shows how the dependence between variable of interest and design variables affects the covariance kernel of the Gaussian limiting law of W_N^H. If compared to Proposition 2 in Conti [17], its main consequence is that, whenever there is some kind of dependence between the design variables (or, equivalently, the sampling weights) and the variable of interest, the empirical process (25) does not converge weakly to a Brownian bridge, but to a Gaussian process with a covariance kernel having a complicated form, depending on the relationships between the character of interest and the design variables. The form of such a relationship is usually unknown.

Before ending the present section we note, *in passing*, that Proposition 1 implies that, with $\mathbb{P}$-probability 1, conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$:

$$\big|\widehat{F}_H(y) - F_N(y)\big| \xrightarrow{p} 0 \quad \text{as } N \to \infty \tag{29}$$

where the symbol $\xrightarrow{p}$ denotes the convergence in probability w.r.t. the sampling design (or better, w.r.t. the sequence of sampling designs in A3). Using the same arguments as the proof of the Glivenko–Cantelli theorem, it is not difficult to prove the following further result.

Proposition 2. *If the sampling design P satisfies assumptions* A1–A6, *with $\mathbb{P}$-probability* 1, *conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\sup_y |\widehat{F}_H(y) - F_N(y)|$ converges to* 0 *in probability w.r.t. the sampling design.*

6. Estimating finite population parameters

The goal of the present section is to study the large sample distribution of estimators of the finite population parameters that are functions of p.d.f. $F_N(\cdot)$. In particular, we concentrate on estimators of the form (21). In a sense, the results of the present section can be viewed as a finite population version of the theory of statistical functionals, that mainly refers to the case of *i.i.d.* observations (cfr. van der Vaart [42], Ch. 20).

The appropriate tool to study asymptotic properties of statistical functionals is the notion of Hadamard-differentiability. Let $\theta(\cdot) : l^\infty[-\infty, +\infty] \to E$ be a map having as domain the normed space $l^\infty[-\infty, +\infty]$ (endowed with the sup-norm), and taking values on an appropriate normed space E with norm $\|\cdot\|_E$. The map $\theta(\cdot)$ is Hadamard-differentiable at F if there exists a continuous linear mapping $\theta'_F : l^\infty[-\infty, +\infty] \to E$ such that

$$\left\| \frac{\theta(F + t h_t) - \theta(F)}{t} - \theta'_F(h) \right\|_E \to 0 \quad \text{as } t \downarrow 0, \text{ for every } h_t \to h. \tag{30}$$

The quantity $\theta'_F(\cdot)$ is the *Hadamard derivative* of θ at F. Let us consider the (sequence of) stochastic process

$$T_N^H = \sqrt{n}\big(\theta(\widehat{F}_H) - \theta(F_N)\big), \quad N \geq 1. \tag{31}$$

In view of Theorem 20.8 in van der Vaart [42] and Proposition 1, the following result holds.

Proposition 3. *Suppose that $\theta(\cdot)$ is (continuously) Hadamard-differentiable at F, with Hadamard derivative $\theta'_F(\cdot)$. Under assumptions* A1–A6, *with $\mathbb{P}$-probability* 1, *conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, the sequence $(T_N^H; N \geq 1)$ converges weakly to $\theta'_F(W^H)$, as N increases.*

Proposition 3 essentially provides, under mild conditions, an asymptotic approximation for the sampling distribution of T_N^H. In particular, if θ is real-valued, since $\theta'_F(\cdot)$ is linear and W^H is a Gaussian process, the law of $\theta'_F(W^H)$ is normal with mean zero and variance

$$\sigma_\theta^2 = \mathbb{E}\big[\theta'_F\big(W^H\big)^2\big]. \tag{32}$$

7. A class of resampling procedure and its basic properties

The main goal of this section is to provide a sound theoretical justification of the two-phase resampling approach described in Section 3. Our argument is of asymptotic nature: the probability distribution of the estimator $\theta(\widehat{F}_H)$ and its approximation based on resampling both converge to the *same* limit. This is actually the main argument in favour of the classical (nonparametric) bootstrap for *i.i.d.* data: cfr., for instance, Bickel and Freedman [7]. The results of the present section can be viewed as an attempt to reconciliate the arguments used in sampling finite populations with those used in classical nonparametric statistics.

The first attempt to define a resampling technique for finite populations based on asymptotic distribution theory is in Chatterjee [14] for simple random sampling, and in Conti and Marella

[18] for general designs. In the latter paper, a technique based on rescaling classical bootstrap is proposed, and its properties are studied. However, two points have to be stressed. The first one is that the technique developed in Conti and Marella [18] is specifically designed to estimate quantiles. The second one is that it is fully justified from an asymptotic point of view only when there are no relationships between π_is (and hence x_is) and y_is. In other words, the rescaled bootstrap proposed in Conti and Marella [18] does not work when the dependence between y_is and x_is cannot be neglected.

In view of the above remarks, in this section a new resampling algorithm for finite population is introduced, that works

(i) for *general* estimators $\theta(\widehat{F}_H)$ of general population parameters $\theta(F_N)$;
(ii) when x_is (*i.e.*, the design variables) and y_is (*i.e.*, the variable of interest) are related by some kind of dependence. No special assumption is made on the relationship between x_is and y_is.

In the sequel, the term *sampling design* P denotes the sampling procedure drawing n units from the "original" population $\mathcal{U}_N$. The *resampling design* P^* is the sampling procedure drawing n units from the predicted (pseudo-)population $\mathcal{U}^*_{N^*}$. Details of the two phases on which the resampling procedure relies are in Sections 7.1, 7.2.

7.1. Phase 1: Pseudo-population

Consider a pseudo-population $\mathcal{U}^*_{N^*}$

$$\left\{\left(N_i^* D_i, y_i, x_i\right); i = 1, \ldots, N\right\}$$

where $N_i^* D_i$ population units are predicted to have y-value equal to y_i and x-value equal to x_i, for each sample unit i. The d.f. of the pseudo-population is equal to

$$F^*_{N^*}(y) = \frac{1}{N^*} \sum_{k=1}^{N^*} I_{(y_k^* \le y)} = \sum_{i=1}^{N} \frac{N_i^*}{N^*} D_i I_{(y_i \le y)}, \quad y \in \mathbb{R} \tag{33}$$

where N^* (3) is the number of pseudo-population units.

As far as the terms N_i^* are concerned, we will make the following assumptions on expectations, variances, covariances w.r.t. P_{pred}.

P1. $E[N_i^* | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N] = \pi_i^{-1} D_i K_{1N}(\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N)$
P2. $V(N_i^* | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N) \le \pi_i^{-1} D_i K_{2N}(\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N)$
P3. $|C(N_i^*, N_h^* | \boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N)| \le \frac{c}{N} \pi_i^{-1} \pi_h^{-1} D_i D_h K_{3N}(\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N)\ i \ne h$

c being a (finite) constant, with

$$K_{1N}(\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N) \to 1 \tag{34}$$

and $K_{jN}(\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N)$, $j = 2, 3$ are bounded in probability, conditionally on $\boldsymbol{D}_N, \boldsymbol{Y}_N, \boldsymbol{X}_N$, as N increases. The symbol $\to$ in (34) denotes convergence in probability w.r.t. $\boldsymbol{D}_N$ and for almost all y_is, x_is.

7.2. Phase 2: Resampling design from the pseudo-population

In phase 2 a sample $\boldsymbol{s}^*$ of size n (the same as the original sample) is selected from the predicted population according to a resampling design P^* satisfying the high entropy condition A5 and with first order inclusion probabilities $\pi_k^* = nx_k^*/\sum_{h=1}^{N^*} x_h^*$. The most obvious choice is to use a resampling design of the same kind as the original sampling design, but with first order inclusion probabilities π_k^*s, even if the use of a resampling design different from the original one could be justified by reduction of the computational burden when N, n are large. The Hájek estimator of the d.f. of the predicted population $F_{N^*}^*(y)$ is equal to

$$\widehat{F}_H^*(y) = \frac{\sum_{k=1}^{N^*} \frac{D_k^*}{\pi_k^*} I_{(y_k^* \le y)}}{\sum_{k=1}^{N^*} \frac{D_k^*}{\pi_k^*}} \tag{35}$$

where $D_k^* = 1$ if the unit k of the predicted population is drawn, and $D_k^* = 0$ otherwise.

Next proposition shows that, in terms of size N^*, the pseudo-population in equivalent to the actual one.

Proposition 4. *Under assumptions* A1–A6, P1–P3, *for almost all* y_is, x_is *values*, *and in probability w.r.t.* $\boldsymbol{D}_N$,

$$\frac{N^*}{N} \to 1 \quad \textit{in probability w.r.t. } P_{\text{pred}} \tag{36}$$

as N *goes to infinity.*

Constructing the pseudo-population and drawing samples from it essentially adds a further "randomness layer" to the whole sampling process. The behaviour of such an additional randomness layer is studied in Proposition 5, where it is shown that sampling from the pseudo-population is asymptotically equivalent to sampling from the original population. The proof of Proposition 5 is fairly similar to the proof of Lemmas 1–5.

Proposition 5. *Under assumptions* A1–A6, P1–P3, *conditionally on* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, *as* N *increases the statements of Lemmas* 1–5 *hold true for the predicted population*, *and for almost all* y_is, x_is *values*, *and in probability w.r.t.* $\boldsymbol{D}_N$ *and* P_{pred}.

The statement "in probability w.r.t. $\boldsymbol{D}_N$" means that the set of $\boldsymbol{D}_N$s values, for which Proposition 5 holds, possesses a probability tending to 1 as N increases.

Define now the "resampled version" of the processes W_N^H (25) and T_N^H (31), namely

$$W_N^{H*} = \left(\sqrt{n}\left(\widehat{F}_H^*(y) - F_{N^*}^*(y)\right), y \in \mathbb{R}\right), \quad N \ge 1; \tag{37}$$

$$T_N^{H*} = \sqrt{n}\left(\theta\left(\widehat{F}_H^*\right) - \theta\left(F_{N^*}^*\right)\right), \quad N \ge 1. \tag{38}$$

Proposition 6 contains the main result of the present section and it can be proved essentially with the same technique as Propositions 1, 3, respectively.

Proposition 6. *Suppose that the sampling design P and the resampling design P^* both satisfy assumptions* A1–A6, *and that conditions* P1–P3 *are fulfilled. Conditionally on* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, $(D_1 N_1^*, \ldots, D_N N_N^*)$, *the following statements hold.*

R1. *The sequence* $(W_N^{H*}; N \geq 1)$ *converges weakly, in* $D[-\infty, +\infty]$ *equipped with the Skorokhod topology, to a Gaussian process* W^H *with zero mean function and covariance kernel* (26).

R2. *If* $\theta(\cdot)$ *is continuously Hadamard differentiable at* F, *then* $(T_N^{H*}; N \geq 1)$ *converges weakly to* $\theta'_F(W^H)$, *as* N *increases.*

In both R1, R2 *weak convergence takes place for a set of* y_i*s,* x_i*s having* $\mathbb{P}$*-probability* 1, *and for a set of* $\boldsymbol{D}_N$*s and* $(N_1^*, \ldots, N_N^*)$ *of probability tending to* 1.

Proposition 6 shows that the resampled process W_N^{H*} (T_N^{H*}) possesses the same limiting law as the "original" process W_N^H (T_N^H) in Proposition 1 (3). In other words, the proposed resampling procedure asymptotically recovers the probability law of $W_N^H(\cdot)$ and $T_N^H(\cdot)$, respectively.

From a technical point of view, Proposition 6 does not require that the resampling design coincides with the original sampling design, as in Holmberg [28], even if this is the most intuitive choice. The essential required conditions are two: (i) the predicted population is constructed as in phase 1; (ii) the first order inclusion probabilities of the resampling design are proportional to the corresponding x_i values, exactly as the original sampling design. Intuitively speaking, this happens because both the original sampling design and the resampling design possess high entropy, and in this case their limiting behaviour essentially depends on the first order inclusion probabilities.

In Proposition 6, the probability distribution of W_N^{H*} (T_N^{H*}) is considered conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, $(D_1 N_1^*, \ldots, D_N N_N^*)$. In other terms, the predicted population is considered as *fixed* (as well as $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$), and the only source of variability is the resampling design from the predicted population. Using Lemmas 1.1, 1.2 in Csörgő and Rosalsky [21], it is possible to see that the same result also holds when one considers the distribution of W_N^{H*} (T_N^{H*}) conditionally on $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$. In this case only $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$ are taken as fixed, and there are *two* sources of variability: (i) the variability of the process generating the predicted population and (ii) the variability of the resampling design from the predicted population. More precisely, the following proposition (that can be proved with the same reasoning as in Csörgő and Rosalsky [21], based on Lemmas 1.1, 1.2 in the above paper) holds true.

Proposition 7. *Suppose the sampling design P and the resampling design P^* satisfy assumptions* A1–A6. *Conditionally on* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, *the following statements hold.*

U1. *The sequence* $(W_N^{H*}; N \geq 1)$ *converges weakly, in* $D[-\infty, +\infty]$ *equipped with the Skorokhod topology, to a Gaussian process* W^H *with zero mean function and covariance kernel* (26).

U2. *If* $\theta(\cdot)$ *is continuously Hadamard differentiable at* F, *then* $(T_N^{H*}; N \geq 1)$ *converges weakly to* $\theta'_F(W^H)$, *as* N *increases.*

In both U1, U2 *weak convergence takes place for a set of* y_i*s,* x_i*s having* $\mathbb{P}$*-probability* 1, *and for a set of* $\boldsymbol{D}_N$*s of probability tending to* 1.

The main consequence of Propositions 6, 7 is that in generating the bootstrap samples two different approaches can be followed:

1.1 *Conditional Approach*: construct a predicted population and generate M bootstrap samples $\boldsymbol{s}^*$ from it;
1.2 *Unconditional approach*: construct M predicted populations and generate one bootstrap sample $\boldsymbol{s}^*$ from each of them.

Clearly, the unconditional approach is more computationally intensive and time consuming than the conditional one.

The basic steps of the resampling procedure are described below. To simplify the notation, in the sequel we will assume that $\theta(\cdot)$ is real-valued, *that is*, we will consider the case of scalar population parameters.

Step 1 Generate M independent bootstrap samples $\boldsymbol{s}^*$ of size n on the basis of the two-phase procedure described above.

Step 2 For each bootstrap sample, compute the corresponding Hájek estimator (35). They will be denoted by $\widehat{F}^*_{H,m}(y)$, $m = 1, \ldots, M$.

Step 3 Compute the corresponding estimates of $\theta(\cdot)$:

$$\widehat{\theta}^*_m = \theta\big(\widehat{F}^*_{H,m}\big); \quad m = 1, \ldots, M.$$

Step 4 Compute the M quantities

$$Z^*_{n,m} = \sqrt{n}\big(\widehat{\theta}^*_m - \theta\big(F^*_{N^*}\big)\big) = \sqrt{n}\big(\theta\big(\widehat{F}^*_{H,m}\big) - \theta\big(F^*_{N^*}\big)\big); \quad m = 1, \ldots, M. \tag{39}$$

Step 5 Compute the variance of (39):

$$\widehat{S}^{2*} = \frac{1}{M-1}\sum_{m=1}^{M}\big(Z^*_{n,m} - \overline{Z}^*_M\big)^2 = \frac{n}{M-1}\sum_{m=1}^{M}\big(\widehat{\theta}^*_m - \overline{\theta}^*_M\big)^2 \tag{40}$$

where

$$\overline{Z}^*_M = \frac{1}{M}\sum_{m=1}^{M} Z^*_{n,m}, \qquad \overline{\theta}^*_M = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}^*_m.$$

Denote further by

$$\widehat{R}^*_{n,M}(z) = \frac{1}{M}\sum_{m=1}^{M} I_{(Z^*_{n,m}\le z)}, \quad z \in \mathbb{R} \tag{41}$$

the empirical distribution function of $Z^*_{n,m}$s, and by

$$\widehat{R}^{*-1}_{n,M}(p) = \inf\big\{z : \widehat{R}^*_{n,M}(z) \ge p\big\}, \quad 0 < p < 1 \tag{42}$$

the corresponding pth quantile.

The empirical d.f. (41) is essentially an approximation of the (resampling) distribution of T_N^{H*} as defined by equation (38). In Proposition 8, it is shown that it converges to the same limit as the d.f. of T_N^{H*}, and that a similar result holds for the quantiles (42).

Proposition 8. *Suppose that assumptions* A1–A6 *are satisfied, let* σ_θ^2 *be defined as in* (32), *let* Φ_{0,σ_θ^2} *be a normal distribution function with expectation* 0 *and variance* σ_θ^2, *and let* $\Phi_{0,\sigma_\theta^2}^{-1}(p)$ *be the* p*-quantile of* Φ_{0,σ_θ^2} (i.e., *the unique solution of* $\Phi_{0,\sigma_\theta^2}(z) = p$), $0 < p < 1$.

For almost all y_i*s,* x_i*s values, and in probability w.r.t.* $\boldsymbol{D}_N$, $(N_1^*, \ldots, N_N^*)$, *conditionally on* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, $(N_1^*, \ldots, N_N^*)$, *the following results hold*:

$$\sup_z \left| \widehat{R}_{n,M}^*(z) - \Phi_{0,\sigma_\theta^2}(z) \right| \xrightarrow{\text{a.s.}} 0; \tag{43}$$

$$\widehat{R}_{n,M}^{*-1}(p) \xrightarrow{\text{a.s.}} \Phi_{0,\sigma_\theta^2}^{-1}(p), \quad \forall 0 < p < 1 \tag{44}$$

as M, N *go to infinity.*

In addition, if the sequence $(Z_m^* - \overline{Z}_M^*)^2$ *is dominated by a r.v.* U *with finite expectation*, that is, $(Z_m^* - \overline{Z}_M^*)^2 \leq U$ *for each* n, N *and* M, *then in probability w.r.t.* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, $(N_1^*, \ldots, N_N^*)$, *conditionally on* $\boldsymbol{y}_N$, $\boldsymbol{x}_N$, $\boldsymbol{D}_N$, $(N_1^*, \ldots, N_N^*)$ *it yields*

$$\widehat{S}^{2*} \to \sigma_\theta^2 \quad \text{as } M, N \to \infty \tag{45}$$

where convergence in (45) *is in probability w.r.t. resampling replications.*

The main consequences of Proposition 8 are two. First of all, *the estimator* $\widehat{S}^{2*}$ *is a consistent estimator of the variance of* $\theta(\widehat{F}_H)$. In the second place, *the confidence intervals*

$$\left[\widehat{\theta}_H - n^{-1/2} R_{n,M}^{*-1}(1-\alpha/2), \widehat{\theta}_H - n^{-1/2} R_{n,M}^{*-1}(\alpha/2)\right] \tag{46}$$

$$\left[\widehat{\theta}_H - n^{-1/2} z_{\alpha/2} \widehat{S}^*, \widehat{\theta}_H + n^{-1/2} z_{\alpha/2} \widehat{S}^*\right] \tag{47}$$

both possess asymptotic confidence level $1 - \alpha$ as N and M increase.

8. Theoretical properties of calibrated pseudo-populations

In view of Proposition 6, all techniques to construct a pseudo-population are asymptotically equivalent, provided that they satisfy conditions P1–P3 of Section 7.1. In this sense, in the present paper a unified approach for resampling based on pseudo-populations is given. However in practical applications, *that is*, for finite n, a crucial aspect that would potentially affect the performance of resampling, is how the pseudo-population is constructed. In the present section, theoretical properties of the (calibrated) pseudo-populations introduced in Section 3 are studied.

8.1. Holmberg pseudo-population

Holmberg pseudo-population has been introduced in Section 3.1. Its size $N^{*\mathrm{HT}}$ is not generally equal to N. However, the ratio $N^{*\mathrm{HT}}/N$ tends in probability to 1 as N, n increase, as it may be easily proved. Furthermore, HT pseudo-population satisfies the regularity conditions P1–P3, and hence the resampling distribution of $\sqrt{n}(\theta(\widehat{F}^*_H) - \theta(F^*_{N^*}))$ tends to the same limit as the sampling distribution of $\sqrt{n}(\theta(\widehat{F}_H) - \theta(F_N))$.

8.2. Multinomial pseudo-population

Multinomial pseudo-population has been introduced in Section 3.2. The r.v. $(N_i^{*\mathrm{MUL}}; i \in \boldsymbol{s})$ possesses a (conditional) multinomial distribution, with moments (4)–(6). Again, conditions P1–P3 are satisfied, so that the resampling distribution of $\sqrt{n}(\theta(\widehat{F}^*_H) - \theta(F^*_{N^*}))$ tends to the same limit as the sampling distribution of $\sqrt{n}(\theta(\widehat{F}_H) - \theta(F_N))$.

8.3. Conditional Poisson pseudo-population

The Conditional Poisson pseudo-population has been introduced in Section 3.3. It satisfies conditions P1–P3, as established in the next proposition.

Proposition 9. *The conditional Poisson pseudo-population satisfies conditions* P1–P3.

As a consequence, the resampling distribution of $\sqrt{n}(\theta(\widehat{F}^*_H) - \theta(F^*_{N^*}))$ tends to the same limit as the sampling distribution of $\sqrt{n}(\theta(\widehat{F}_H) - \theta(F_N))$.

8.4. Double-calibrated pseudo-population

Double-Calibrated pseudo-population has been introduced in Section 3.4. The main result is in the next proposition.

Proposition 10. *The calibrated pseudo-population with replicates* $N_i^{*\mathrm{DCal}}$ *that solves the optimization problem* (9) *possesses the following property*:

$$\frac{N_i^{*\mathrm{DCal}}}{N_i^{*\mathrm{CPP}}} \xrightarrow{p} 1 \quad \text{as } n, N \to \infty. \tag{48}$$

Intuitively speaking, Proposition 10 tells us that as N, n increase, the solution of the optimization problem (9) tends to coincide with $N_i^{*\mathrm{CPP}}$. Hence, for "very large" population and sample size, $N_i^{*\mathrm{CPP}}$s can be taken as a good approximation of the actual solution of the optimization problem (9). Of course, this is only an asymptotic result, and for the use of "not too large" n, N, the use of $N_i^{*\mathrm{DCal}}$ instead of $N_i^{*\mathrm{CPP}}$ could produce considerably different results in the resampling procedure.

8.5. Hot-deck pseudo-population

Hot-deck pseudo-population has been introduced in Section 3.5. It satisfies conditions P1–P3, as established in the next proposition.

Proposition 11. *If the pseudo-population is constructed* via *hot-deck imputation of ys values, then, as n, N increase, the resampling distribution of* $\sqrt{n}(\theta(\widehat{F}_H^*) - \theta(F_{N^*}^*))$ *tends to the same limit as the sampling distribution of* $\sqrt{n}(\theta(\widehat{F}_H) - \theta(F_N))$.

9. Simulation study

Main goal of the simulation is to empirically evaluate the effects that different choices for constructing the pseudo-population $\mathcal{U}_{N^*}^*$ (where resampling is actually performed) may have upon the accuracy of the resulting inference in practical applications. The simulation has been designed by focusing on three key points:

(a) exploration of small to moderate n and N in order to highlight differences due to finite sizes as well as to evaluate the asymptotic approximations provided in the first part of the present paper;
(b) analysis of specific features of the pseudo-population $\mathcal{U}_{N^*}^*$ due to different construction choices;
(c) investigation of the statistical properties of the final estimates provided by resampling from different pseudo-populations.

The simulated scenarios, parameters and estimators are summarized in Table 1.

In addition to the five strategies proposed in Section 3, the direct bootstrap (Antal and Tillé [1]) is also considered in the simulation, since it is a recent competitor based on a non-predictive resampling approach. For the sake of comparability, the variates Y, X have been simulated under the same model as in Antal and Tillé [1]. In more details, a finite population of size N was generated from the model

$$y_i = \left(\beta_0 + \beta_1 x_i^{1.2} + \sigma\epsilon_i\right)^2 + c \tag{49}$$

Table 1. Simulated scenarios, population parameters and estimators

Scenarios		
$N = 200$, **400**	$n = (0.2N) = 40$, **80**	
correlation between Y and X $\simeq 0.8$		
Parameters	*Hájek Estimators*	*HT Estimator*
$\overline{Y}_N = \sum_{i=1}^N y_i/N$	$\widehat{\overline{Y}}_H = \sum_{i=1}^N D_i\pi_i^{-1}y_i / \sum_{i=1}^N D_i\pi_i^{-1}$	$\widehat{\overline{Y}}_{\mathrm{HT}} = N^{-1}\sum_{i=1}^N D_i\pi_i^{-1}y_i$
$Q_N(p) = \inf\{y : F_N(y) \geq p\}$ with $p = 0.5, 0.75$	$\widehat{Q}_H(p) = \inf\{y : \widehat{F}_H(y) \geq p\}$	

where $x_i = |j_i|$ and $j_i \sim N(0,7)$, $\epsilon_i \sim N(0,1)$ and $\sigma = 15$. The model regression parameters are $\beta_0 = 12.5$, $\beta_1 = 3$ and $c = 4000$. As far as the inclusion probabilities are concerned, they are taken proportional to the value of a variable Z, generated from the equation $Z = Y^{0.2}W$ where W has a lognormal distribution ($\ln N(\mu, \sigma^2)$) with parameters $\mu = 0$ and $\sigma^2 = 0.025$.

Samples have been simulated under two different fixed size πps designs of increasing entropy: Pareto sampling and (normalized) conditional Poisson sampling (CPS for short), this latter already mentioned in Section 4 as a maximum entropy design. Notice that Pareto design is high entropy, although not yet proved asymptotically maximum entropy; however it is heuristically recognized to be very close to the asymptotically maximum entropy Rao–Sampford design (Bondesson *et al.* [9], Lundqvist [32]). Moreover, unlike the CPS design, the Pareto sampling is very simple to implement, and can be used in simple acceptance-rejection rules to produce CPS samples with a significant reduction of computational burden. Simulation has been implemented partly in *Mathematica* code and partly in the R environment. 1000 Monte Carlo (MC) runs, simulating the sample space, have been combined with $M = 1000$ resampling runs from each generated sample. The MC error deriving from these choices has been controlled via the empirical bias of the (unbiased) Horvitz–Thompson estimator $\hat{\bar{Y}}_{\mathrm{HT}}$, and it has been kept under 1% (relative to the true population mean $\bar{Y}$).

Simulation results are gathered in Tables 2–5 where the simulated methods to construct the pseudo-population are indicated by the following acronyms: HT illustrated in Section 3.1; MUL for the Multinomial pseudo-population in 3.2; CPP for the conditional Poisson pseudo-population in 3.3; DCal for the double-calibrated pseudo-population in 3.4; HD for the hot-deck pseudo-population in 3.5; and Dir for the direct bootstrap by Antal and Tillé [1].

Results in Table 2 offer indications about the ability of the pseudo-population $\mathcal{U}^*_{N^*}$ as a *predictor* of the actual population $\mathcal{U}_N$, according to key point (b) above. Except for the direct bootstrap involving no pseudo-population, it has been checked in two respects: (i) the pseudo-population size N^* and mean of the auxiliary variable $\bar{X}^*$ as *predictors* of (known) population N and $\bar{X}_N$ respectively, as measured via empirical (relative) bias $\mathrm{RB}[N^*; N] = 100 \times [E_{\mathrm{MC}}(N^*) - N]/N$

Table 2. $\mathcal{U}^*_{N^*}$ as a predictor of $\mathcal{U}_N$ ($N = 200$, **400**)

	$\mathrm{RB}[\bar{X}^*; \bar{X}_N]$		$\mathrm{RB}[N^*; N]$		$\mathrm{Sup}_{\mathrm{MC}}\lvert F^*_{N^*}(y) - F_N(y)\rvert$	
PARETO sampling design						
HT	0.03	**0.04**	−0.44	**0.38**	0.87	**0.51**
MUL	5.46	**3.39**	0	**0**	0.93	**0.54**
CPP	5.46	**3.38**	0	**0**	0.88	**0.52**
DCal	0.02	**−0.02**	0.003	**−0.01**	0.55	**0.46**
HD	0	**0**	0	**0**	0.47	**0.37**
CPS sampling design						
HT	−0.02	**0.02**	−1.05	**−1.40**	0.50	**0.52**
MUL	5.06	**3.39**	0	**0**	0.53	**0.55**
CPP	5.04	**3.88**	0	**0**	0.51	**0.52**
DCal	0.06	**−0.04**	0.04	**−0.04**	0.46	**0.47**
HD	0	**0**	0	**0**	0.48	**0.33**

Table 3. 95% Resampling CI – *percentile* method ($N = 200$, **400**)

	$\hat{\bar{Y}}_H$				$\hat{Q}_H(0.5)$				$\hat{Q}_H(0.75)$			
	EC		*AL*		*EC*		*AL*		*EC*		*AL*	
PARETO												
HT	0.89	**0.90**	0.23	**0.17**	0.88	**0.91**	0.33	**0.22**	0.91	**0.93**	0.37	**0.28**
MUL	0.87	**0.89**	0.23	**0.17**	0.67	**0.79**	0.33	**0.14**	0.82	**0.79**	0.38	**0.28**
CPP	0.89	**0.90**	0.23	**0.17**	0.89	**0.92**	0.33	**0.22**	0.92	**0.94**	0.38	**0.29**
DCal	0.95	**0.95**	0.24	**0.18**	0.95	**0.97**	0.33	**0.23**	0.95	**0.95**	0.39	**0.30**
HD	0.97	**0.98**	0.27	**0.20**	0.95	**0.95**	0.36	**0.26**	0.99	**0.99**	0.43	**0.33**
CPS												
HT	0.90	**0.91**	0.24	**0.17**	0.91	**0.92**	0.33	**0.22**	0.90	**0.93**	0.38	**0.29**
MUL	0.89	**0.92**	0.24	**0.17**	0.73	**0.81**	0.34	**0.22**	0.82	**0.79**	0.39	**0.29**
CPP	0.90	**0.90**	0.24	**0.17**	0.91	**0.92**	0.34	**0.22**	0.91	**0.94**	0.38	**0.29**
DCal	0.96	**0.96**	0.25	**0.18**	0.98	**0.97**	0.34	**0.23**	0.95	**0.94**	0.40	**0.30**
HD	0.98	**0.98**	0.27	**0.20**	0.96	**0.95**	0.37	**0.26**	0.99	**0.99**	0.44	**0.33**

(where E_{MC} indicates the average over all the Monte Carlo runs and $\mathrm{RB}[\bar{X}^*; \bar{X}_N]$ follows accordingly); and (ii) how able the pseudo-population is to reproduce the actual p.d.f. as measured by the maximal MC value of the Kolmogorov statistic $\max_{\mathrm{MC}} \sup_y |F^*_{N*}(y) - F_N(y)|$, $y \in \mathbb{R}$.

A clear connection appears between the conservation of both N and $\bar{X}$ and the ability of reproducing the entire population d.f.: HD and DCal pseudo-populations emerge as the best performers, uniformly in all the simulated scenarios. Also, this reflects on the ability of the resampling algorithm based on such pseudo-populations, to reproduce the estimator distribution.

According to key point (c) above, both kinds of confidence intervals (CI) illustrated in Section 7 have been simulated. Table 3 concerns CI (46) which basically correspond to *bootstrap percentile* method, and Table 4 refers to CI (47). Performances at (nominal) confidence level 95% has been investigated *via* empirical coverage (EC), with respect to the true population parameter, and average length (AL). Notice that although the *percentile* method is the crudest available for producing CI via resampling, we rate it appropriate for the goals of the present simulation because it allows the evaluation of the ability of the resampling algorithm to produce p-values, and ultimately to reproduce the estimator sampling distribution particularly in its tails. In Table 3 all the methods investigated for constructing $\mathcal{U}^*_{N*}$ provide acceptable levels of empirical coverage based on the 0.025 and 0.975 percentiles of the resampling distribution. Moreover, they all tend to improve for increasing sizes N and n, as expected according to asymptotic results in Section 7. However HD and DCal, which provide the best predictor of $\mathcal{U}_N$, also give the best coverage probabilities, uniformly in all scenarios simulated for both linear and non linear estimators. Notably, HD shows the largest average lengths in addition to the largest empirical coverages, which suggests a tendency to supply conservative CI.

A similar behaviour can be observed in Table 4, although the resampling plays here a minor role, limited to the (point) bootstrap estimate (40) for the estimator variance then coupled with standard normal distribution percentiles, also named bootstrap-t CI. Notice that this is also the method for interval estimation suggested for the *non-predictive* direct bootstrap. However,

Table 4. 95% Standard Normal CI with resampling variance estimate ($N = 200, \mathbf{400}$)

	$\hat{\bar{Y}}_H$				$\hat{Q}_H(0.5)$				$\hat{Q}_H(0.75)$			
	EC		*AL*		*EC*		*AL*		*EC*		*AL*	
PARETO												
HT	0.90	**0.91**	0.24	**0.17**	0.90	**0.91**	0.36	**0.24**	0.93	**0.91**	0.40	**0.30**
MUL	0.90	**0.91**	0.24	**0.17**	0.89	**0.92**	0.36	**0.24**	0.92	**0.91**	0.41	**0.30**
CPP	0.91	**0.92**	0.24	**0.17**	0.89	**0.92**	0.36	**0.24**	0.93	**0.92**	0.40	**0.30**
DCal	0.84	**0.86**	0.25	**0.18**	0.85	**0.89**	0.38	**0.25**	0.88	**0.90**	0.43	**0.33**
HD	0.91	**0.93**	0.27	**0.20**	0.92	**0.94**	0.40	**0.27**	0.95	**0.96**	0.44	**0.34**
Dir	0.89	**0.90**	0.22	**0.16**	0.86	**0.87**	0.32	**0.21**	0.92	**0.90**	0.38	**0.28**
CPS												
HT	0.91	**0.91**	0.24	**0.17**	0.89	**0.90**	0.37	**0.24**	0.92	**0.90**	0.40	**0.30**
MUL	0.90	**0.92**	0.24	**0.17**	0.89	**0.92**	0.38	**0.24**	0.90	**0.90**	0.41	**0.31**
CPP	0.91	**0.92**	0.24	**0.17**	0.90	**0.91**	0.37	**0.24**	0.92	**0.91**	0.40	**0.31**
DCal	0.85	**0.87**	0.25	**0.19**	0.87	**0.87**	0.39	**0.25**	0.89	**0.89**	0.44	**0.33**
HD	0.94	**0.95**	0.27	**0.20**	0.90	**0.93**	0.40	**0.27**	0.97	**0.95**	0.45	**0.34**
Dir	0.90	**0.90**	0.23	**0.16**	0.85	**0.88**	0.33	**0.21**	0.92	**0.88**	0.38	**0.28**

Dir exhibits lower empirical coverage probabilities than the *predictive* pseudo-population based methods, seemingly due to systematic smaller lengths. The notable exception of DCal may be explained by its weaker ability to produce accurate point bootstrap estimates than the other *predictive* methods simulated. Still HD emerges as the best performer for uniformly giving the larger empirical coverages in all scenarios simulated and for maintaining its conservative peculiarity.

Finally, a popular property of the classic *i.i.d.* Efron's bootstrap has been investigated, *that is*, the ability of the resampled distribution of an estimator of the population mean to match the (original) sample mean as its empirical first moment. Such property, dubbed *bootstrap unbiasedness*, has been measured by the (percentage) relative bias $\mathrm{RB}[\hat{\theta}^*_m; \hat{\theta}] = 100 \times E_{\mathrm{MC}}\{[E^*(\hat{\theta}^*_m) - \hat{\theta}]/\hat{\theta}\}$ where E^* indicates the empirical average over the M resampling runs and by taking $\hat{\theta} = \bar{Y}$ and $\hat{\theta}^*_m, m = 1 \cdots M$ as its resampled distribution. Table 5 reports simulation results with respect to both Horvitz–Thompson and Hájek estimation of population mean. Empirical evidence highlights that HT and Dir perform better under the conventional Horvitz–Thompson estimation of linear parameters, as it is expected by their construction.

As a final remark concerning the actual implementation of specific algorithms, note that all the simulated populations have been checked to ensure $\pi_i < 1$, $i = 1, \ldots, N$. However, for MUL it may still occur $\pi^*_k \geq 1$ for one or more (sampled) unit k included in the pseudo-population. This empirically appears to be often the case as the number of MC runs increases. As a consequence, an *ad hoc* routine has to be implemented on top of the resampling algorithm, aiming at including such units in each bootstrap sample and sequentially recomputing the resampling inclusion probability until they are all strictly smaller than 1, and by simultaneously reducing the (re)sample size accordingly (see, for instance, Tillé [41] for details).

Table 5. Bootstrap-unbiasedness ($N = 200, \mathbf{400}$)

	PARETO				CPS			
	$\mathrm{RB}[\hat{\hat{Y}}^*_{\mathrm{HT}} - \hat{\hat{Y}}_{\mathrm{HT}}]$		$\mathrm{RB}[\hat{\hat{Y}}^*_{H} - \hat{\hat{Y}}_{H}]$		$\mathrm{RB}[\hat{\hat{Y}}^*_{\mathrm{HT}} - \hat{\hat{Y}}_{\mathrm{HT}}]$		$\mathrm{RB}[\hat{\hat{Y}}^*_{H} - \hat{\hat{Y}}_{H}]$	
HT	0.06	**−0.16**	0.87	**0.72**	−0.16	**−0.13**	0.97	**0.63**
MUL	5.57	**3.11**	0.92	**0.65**	4.96	**3.74**	1.07	**0.65**
CPP	5.46	**3.15**	0.84	**0.70**	4.84	**3.73**	1.01	**0.64**
DCal	1.66	**1.12**	−0.33	**0.20**	1.38	**1.36**	−0.34	**−0.11**
HD	3.17	**2.07**	0.35	**0.59**	2.55	**2.27**	0.34	**0.28**
Dir	0.01	**−0.01**	0.70	**0.41**	−0.02	**0.01**	0.68	**0.41**

10. Conclusions

In this paper, a new class of resampling methods applying to non-*i.i.d.* finite population sampling is proposed under a principled *predictive* approach. The proposed resampling unifies any method based on pseudo-populations, i.e. according to the *plug-in* principle upon which the original Efron's bootstrap is based. A large sample theory is derived for the predictive resampling, in the Hájek finite population asymptotic setup, and according to the classical asymptotics for *i.i.d.* bootstrap by Bickel and Freedman [7]. It is also proved that all techniques producing the pseudo-population are asymptotically equivalent, under mild regularity conditions.

In addition, five strategies for constructing the pseudo-population have been illustrated. Two of them go back to results already appeared in the literature and the remaining three are new proposals with improved performance, as shown in the simulation study. Empirical evidence confirms that how to construct the pseudo-population is a crucial choice for small to moderate population and sample sizes, under general sampling designs such as πps designs. As a general recommendation such choice should be guided by enforcing the ability of the pseudo-population to be a *good predictor* of the actual population. The simulation study indicates the pseudo-population based on hot-deck imputation (HD) as the soundest method, provided that auxiliary x_is values are available for all population units. When x_is are known only for sample units, as it might be the case in applications, good results are offered by a pseudo-population calibrated w.r.t. both the population size and the mean (total) of the auxiliary variable (DCal), when combined with percentile confidence intervals.

Supplementary Material

Supplement to "A unified principled framework for resampling based on pseudo-populations: Asymptotic theory" (DOI: 10.3150/19-BEJ1138SUPP; .pdf). Supplementary information.

References

[1] Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *J. Amer. Statist. Assoc.* **106** 534–543. MR2847968 https://doi.org/10.1198/jasa.2011.tm09767

[2] Beaumont, J.-F. and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *Int. Stat. Rev.* **80** 127–148. MR2990349 https://doi.org/10.1111/j.1751-5823.2011.00166.x

[3] Berger, Y.G. (1998). Rate of convergence to normal distribution for the Horvitz–Thompson estimator. *J. Statist. Plann. Inference* **67** 209–226. MR1624693 https://doi.org/10.1016/S0378-3758(97)00107-9

[4] Berger, Y.G. (2005). Variance estimation with Chao's sampling scheme. *J. Statist. Plann. Inference* **127** 253–277. MR2103037 https://doi.org/10.1016/j.jspi.2003.08.014

[5] Berger, Y.G. (2011). Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan J. Statist.* **27** 407–426. MR2919728

[6] Bertail, P., Chautru, E. and Clémençon, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** 97–111. MR3619696 https://doi.org/10.1111/sjos.12243

[7] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103

[8] Boistard, H., Lopuhaä, H.P. and Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. *Ann. Statist.* **45** 1728–1758. MR3670194 https://doi.org/10.1214/16-AOS1507

[9] Bondesson, L., Traat, I. and Lundqvist, A. (2006). Pareto sampling versus Sampford and conditional Poisson sampling. *Scand. J. Stat.* **33** 699–720. MR2300911 https://doi.org/10.1111/j.1467-9469.2006.00497.x

[10] Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. *J. Amer. Statist. Assoc.* **89** 1282–1289. MR1310222

[11] Brewer, K.R.W. and Donadio, M.E. (2003). The high entropy variance of the Horvitz–Thompson estimator. *Surv. Methodol.* **29** 189–196.

[12] Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York–London–Sydney: Wiley Interscience. MR0652527

[13] Chao, M.T. and Lo, S.-H. (1985). A bootstrap method for finite population. *Sankhya, Ser. A* **47** 399–405. MR0863733

[14] Chatterjee, A. (2011). Asymptotic properties of sample quantiles from a finite population. *Ann. Inst. Statist. Math.* **63** 157–179. MR2748939 https://doi.org/10.1007/s10463-008-0210-4

[15] Chauvet, G. (2007). Méthodes de bootstrap en population finie. Ph.D. Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Universioté de Rennes 2.

[16] Chen, X.-H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika* **81** 457–469. MR1311090 https://doi.org/10.1093/biomet/81.3.457

[17] Conti, P.L. (2014). On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B* **76** 234–259. MR3302272 https://doi.org/10.1007/s13571-014-0083-x

[18] Conti, P.L. and Marella, D. (2015). Inference for quantiles of a finite population: Asymptotic versus resampling results. *Scand. J. Stat.* **42** 545–561. MR3345121 https://doi.org/10.1111/sjos.12122

[19] Conti, P.L., Marella, D., Mecatti, F. and Andreis, F. (2020). Supplement to "A unified principled framework for resampling based on pseudo-populations: Asymptotic theory." https://doi.org/10.3150/19-BEJ1138SUPP.

[20] Conti, P.L., Marella, D. and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *J. Amer. Statist. Assoc.* **111** 1715–1725. MR3601730 https://doi.org/10.1080/01621459.2015.1112803

[21] Csörgő, S. and Rosalsky, A. (2003). A survey of limit laws for bootstrapped sums. *Int. J. Math. Math. Sci.* **45** 2835–2861. MR2005871 https://doi.org/10.1155/S0161171203301437
[22] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
[23] Efron, B. (2003). Second thoughts on the bootstrap. *Statist. Sci.* **18** 135–140. MR2026075 https://doi.org/10.1214/ss/1063994968
[24] Grafström, A. (2010). Entropy of unequal probability sampling designs. *Stat. Methodol.* **7** 84–97. MR2591712 https://doi.org/10.1016/j.stamet.2009.10.005
[25] Gross, S.T. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Reasearch Methods*. American Statistical Association 181–184.
[26] Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Ann. Math. Stat.* **35** 1491–1523. MR0178555 https://doi.org/10.1214/aoms/1177700375
[27] Hájek, J. (1981). *Sampling from a Finite Population*. *Statistics*: *Textbooks and Monographs* **37**. New York: Dekker. MR0627744
[28] Holmberg, A. (1998). A bootstrap approach to probability proportional-to-size sampling. In *Proceedings of the ASA Section on Survey Research Methods* 378–383.
[29] Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96. MR0648029
[30] Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statist. Sci.* **18** 199–210. MR2019788 https://doi.org/10.1214/ss/1063994975
[31] Lahiri, S.N. (2003). *Resampling Methods for Dependent Data*. *Springer Series in Statistics*. New York: Springer. MR2001447 https://doi.org/10.1007/978-1-4757-3803-2
[32] Lundqvist, A. (2007). On the distance between some πps sampling designs. *Acta Appl. Math.* **97** 79–97. MR2329721 https://doi.org/10.1007/s10440-007-9134-x
[33] Marella, V. and Vicard, P. (2017). Structural learning for complex survey data. *Cladag 2017. 11th Scientific Meeting of the Classificaton and Data Analysis Group. Book of Short Papers*. (ISBN 978-88-9945971-0).
[34] Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Stat. Surv.* **10** 1–52. MR3476140 https://doi.org/10.1214/16-SS113
[35] McCarthy, P.J. and Snowden, C.B. (1985). The bootstrap and finite population sampling. In *Vital and Health Statistics* **95**(2) 1–23. Washington, DC: Public Heath Service Publication, U.S. Government Printing.
[36] Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Int. Stat. Rev.* **61** 317–337.
[37] Ranalli, M.G. and 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* 4088–4099.
[38] Rao, J.N.K. and Wu, C.-F.J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241. MR0941020
[39] Sitter, R.R. (1992). A resampling procedure for complex survey data. *J. Amer. Statist. Assoc.* **87** 755–765. MR1185197
[40] Sverchkov, M. and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Surv. Methodol.* **30** 79–92.
[41] Tillé, Y. (2006). *Sampling Algorithms*. *Springer Series in Statistics*. New York: Springer. MR2225036
[42] van der Vaart, A.W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Received December 2017 and revised May 2019