Resampling under Complex Sampling Designs: Roots, Development and the Way Forward

Pier Luigi Conti 1,* and Fulvia Mecatti 2

1 Dipartimento di Scienze Statistiche, Sapienza Università di Roma, P.le Aldo Moro, 5, 00185 Roma, Italy
2 Dipartimento di Sociologia e Ricerca Sociale, Università di Milano-Bicocca, Via Bicocca Degli Arcimboldi, 8, 20126 Milano, Italy; fulvia.mecatti@unimib.it
* Correspondence: pierluigi.conti@uniroma1.it

Abstract: In the present paper, resampling for finite populations under an iid sampling design is reviewed. Our attention is mainly focused on pseudo-population-based resampling due to its properties. A principled appraisal of the main theoretical foundations and results is given and discussed, together with important computational aspects. Finally, a discussion on open problems and research perspectives is provided.

Keywords: resampling; bootstrap; pseudo-population; asymptotics; empirical processes

1. Introduction
1.1. Generalities

Resampling methods have a long and honorable history, going back at least to the seminal paper by [1]. Survey data are an ideal context to use resampling methods to approximate the sampling distribution of statistics, due to both (i) a generally large sample size and (ii) data of typically good quality.

The present paper does not aim at providing a complete review of resampling methods in sampling statistics; the interested reader is referred, for instance, to [2]. We mainly focus on a special class of resampling methods—namely those based on pseudo-populations. There are several reasons to support this restriction. First of all, they may be viewed, in many respects, as the “natural” extension of classical Efron’s bootstrap to sampling finite populations, in both descriptive and analytic inference (i.e., inference on finite population and superpopulation parameters, respectively).

In the second place, in our knowledge, they are the only methods with a rigorous asymptotic justification in terms of weak convergence of empirical processes, allowing results not only for linear estimators but also for non-linear ones (under suitable differentiability conditions).

In extreme synthesis, virtually all resampling methodologies used in sampling from finite populations are based on the idea of accounting for the effect of the sampling design. As it will be seen in the sequel, the main effect of the sampling design is that data cannot be generally assumed independent and identically distributed (i.i.d.). A large portion of the literature on resampling from finite populations focuses on estimating the variance of estimators. The main approaches are essentially the ad hoc approach and plug in approach.

The basic idea of the ad hoc approach consists in maintaining Efron’s bootstrap as a resampling procedure but in properly rescaling data in order to account for the dependence among units. This approach is used, among others, in [3,4], where the re-sampled data produced by the “usual” i.i.d. bootstrap are properly rescaled, as well as in [5,6]; cfr. also the review in [7]. In [8] a “rescaled bootstrap process” based on asymptotic arguments is proposed. Among the ad hoc approaches, we also classify [9] (based on a rescaling of weights) and the “direct bootstrap” by [10].
Almost all ad hoc resampling techniques are based on the same justification: in the case of linear statistics, the first two moments of the resampled statistic should match (at least approximately) the corresponding estimators; cfr., among the others, [10]. Cfr. also [9], where an analysis in terms of the first three moments is performed for Poisson sampling.

Plug-in approaches, which are considered in the present paper, 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 population. Then, bootstrap samples are drawn from such a pseudo-population according to some appropriate resampling design; cfr. [11–15] as well as [2].

Before entering the subject of resampling, it seems appropriate to give a formal setting for both descriptive and analytic inference.

1.2. Superpopulation Model and Sampling Design: Basic Aspects

Consider a finite population \( \mathcal{U}_N \) of \( N \) units. If \( Y \) denotes the character of interest, let \( y_i \) be the value \( Y \) for unit \( i (= 1, \ldots, N) \). Each \( y_i \) value is assumed to be a realization of a random variable (r.v.) \( Y_i \). The \( N \)-variate r.v. \( Y_N = (Y_1, \ldots, Y_N) \) is the superpopulation. In addition, for every population unit \( J \) further r.v.s, playing the role of auxiliary variables, \( \{T_{ij}, i = 1, \ldots, N\} \) are defined, where \( T_{ij} \) is the value of the \( j \)-th auxiliary variable \( (j = 1, \ldots, J) \) for unit \( i (= 1, \ldots, N) \). The symbol \( T_{N,i} \) will be used, when necessary, to denote the \( N \times J \) matrix of elements \( T_{ij} \). Auxiliary variables play a preeminent role in constructing the sampling design, and, for this reason, they will be called design variables.

For the sake of simplicity, in the sequel, the \((J + 1)\)-dimensional random vectors \((Y_i, T_{i1}, \ldots, T_{ij})\) are assumed to be independent and identically distributed (i.i.d.). They can be thought as the first \( N \) elements of a sequence \( \{(Y_i, T_{i1}, \ldots, T_{ij}); i \geq 1\} \), existing on a probability space \((\Omega_A, \mathcal{A}, P^N)\), where, due to the i.i.d. assumption, \( P^N \) is the product measure of identical copies of a single \( P_\Omega \). The symbols \( E_\Omega, V_\Omega, C_\xi \) denote the corresponding operators of expectation, variance and covariance, respectively.

To define a general sampling design, including both “with replacement” and “without replacement” cases, for each unit \( i \in \mathcal{U}_N \), we consider a discrete random variable (r.v.) \( D_i \) taking values \( 0, 1, \ldots, K_i \) and representing the multiplicity of unit \( i \) within the sample, namely the number of times unit \( i \) appears in the selected sample. The sample membership indicator of unit \( i \) is defined as \( I_i = \min(1, D_i) \). A sampling design is without replacement if \( S_i = \{0, 1\} \) for each unit \( i \), namely if \( D_i = I_i \) for each \( i = 1, \ldots, N \).

A sampling design is essentially the “probabilistic rule” according to which a sample is selected from a finite population, given the values \( y_1, \ldots, y_N \) (and given the values of the design variables, as well). Generally speaking, specifying the sampling design is equivalent to specify the joint distribution of the random vector r.v. \( D_N = (D_1, \ldots, D_N) \). Such a joint distribution will be denoted in the sequel by \( P_p \). It may either depend or not depend on \( y_1, \ldots, y_N \). A sampling design that does not depend on \( y_i \) is non-informative.

In the sequel, a short formal description of sampling designs, based on probability and measure theory, is provided. On first reading, this part can be omitted without affecting the understanding of the main points of the present paper.

Let \( S_i \) be the set \( \{0, 1, \ldots, K_i\} \). In general, the r.v. \( D_N \) is defined on the probability space \((\prod_{i=1}^N S_i, \mathcal{P}(\prod_{i=1}^N S_i), P_{P,N})\), where \( \mathcal{P}(\prod_{i=1}^N S_i) \) is the power set of \( \prod_{i=1}^N S_i \), and \( P_{P,N} \) possesses the following two properties.

(a) \( P_{P,N}(\cdot, Y_N, T_{N,i}) \) is a probability measure on \( (\prod_{i=1}^N S_i, \mathcal{P}(\prod_{i=1}^N S_i)) \) for every \((Y_N, T_{N,i})\) in \( \mathbb{R}^N \times \mathbb{R}^N \).

(b) \( P_{P,N}(B, Y_N, T_{N,i}) \) is a Borel-measurable function of \((Y_N, T_{N,i})\) for every \( B \in \mathcal{P}(\prod_{i=1}^N S_i) \).

The main restriction that we will consider on the sampling design is that it is non-informative, namely

\[ P_{P,N}(\cdot, Y_N, T_{N,i}) = P_{P,N}(\cdot, T_{N,i}) \]
Intuitively speaking, the above relationship means that the probability measure $P_{P,N}$ does not depend on the values of the study variable, $Y_i's$, but only on the design variables. Moreover, $P_{P,N}(\cdot, T_{N,J})$ can be interpreted as the probability measure corresponding to the sampling design conditionally on the design variates.

On the basis of the above elements, a probability space $(\Omega', A', P')$ is defined, where $\Omega' = \Omega \times (\prod_{i=1}^{N} S_i)$, $A' = A \otimes \mathcal{P}(\prod_{i=1}^{N} S_i)$, and

$$P'(A \times B) = \int_{A} P_{P,N}(B, T_{N,J})d\xi.$$ 

To simplify the notation, in the sequel, we denote by $P(P(\cdot))$ the probability distribution of the r.v.s $D_N$, given the values of the design variables ($P(D_N \in B) = P_{P,N}(B)$ for every $B \in \mathcal{P}([0, 1]^N)$) and by $E_P, V_P$ the corresponding operators of expectation, variance, respectively. In particular, the expectations $\pi_i = E_P[I_i]$ and $\pi_{ij} = E_P[I_i I_j]$ are the first and second order inclusion probabilities, respectively. The suffix $P$ denotes the sampling design used to select population units. The (effective) sample size is $n_s = D_1 + \cdots + D_N$ ($v_s = I_1 + \cdots + I_N$).

1.3. Descriptive and Analytic Inference

For the sake of simplicity, let us assume that $Y_1, \ldots, Y_N$ are i.i.d. r.v., with common d.f. $F_\xi$. A superpopulation parameter is a functional (not necessarily real-valued)

$$\theta_\xi = \theta(F_\xi).$$

The simplest example of superpopulation parameter is the expected value

$$\mu = \int_{-\infty}^{+\infty} y dF_\xi(y);$$

however, many other parameters could be of interest.

The finite population distribution function (f.p.d.f., for short) is defined as

$$F_N(y) = \frac{1}{N} \sum_{i=1}^{N} I_{(-\infty, y]}(y_i)$$

A finite population parameter is a functional

$$\theta_N = \theta(F_N).$$

The simplest example is of course the finite population mean:

$$\overline{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i = \int_{-\infty}^{+\infty} y dF_N(y).$$

We note in passim that a finite population parameter $\theta_N$ is a r.v., with probability distribution depending on that of the superpopulation.

Finite population and superpopulation parameters are essentially different in nature, because finite population parameters are observable (it is sufficient to take a census), while superpopulation parameters are not.

The term descriptive inference refers to statistical inference on finite population parameters. On the other hand, the term analytic inference refers to statistical inference on superpopulation parameters.
2. From Efron’s iid Bootstrap to Pseudo-Population Based Resampling

2.1. Efron’s Bootstrap: A Few Basic Aspects

Suppose a sample \( s \) of \( n \) units is drawn from the population \( \mathcal{U}_N \), according to simple random sampling with replacement (srswr) of size \( n \). In practice, \( n \) independent draws are performed, and at each draw, the \( N \) population units have the same probability of being selected. As a consequence, the \( n \) units within sample \( s \) are not necessarily distinct, and the r.v. \( D_N \) has a multinomial distribution with the parameters \( n \) and \( 1/N, \ldots, 1/N \). If \( Y_s = (Y_i; i \in s) \) is the \( n \)-variate r.v. corresponding the our \( n \) sampling observations, then the following two results hold:

- Conditionally on \( Y_N = y_N \), the r.v.s in \( Y_s \) are i.i.d. with common d.f. \( F_N(y) \), the finite population d.f.
- Unconditionally, the r.v.s in \( Y_s \) are i.i.d. with common d.f. \( F_{\mathcal{U}_N}(y) = \mathbb{P}_s((-\infty, y]) \).

In this case, the sampling design does not play any role because the sampling distribution of observations in \( Y_s \) reproduces, both conditionally and unconditionally, the population distribution function.

As a “natural” estimate of the population d.f., it is customary to take the empirical distribution function (e.d.f.):

\[
F_n(y) = \frac{1}{n} \sum_{i \in s} I_{(-\infty, y]}(Y_i) = \frac{1}{n} \sum_{i=1}^N D_i I_{(-\infty, y]}(Y_i). \tag{3}
\]

The e.d.f. (3) is an unbiased estimator of both \( F_N \) and \( F_{\mathcal{U}_N} \).

If the interest is in estimating parameters of the form (1) or (2), then intuition suggests to resort to the statistical functional:

\[
\theta_n = \theta(F_n). \tag{4}
\]

The idea behind Efron’s bootstrap is simple but powerful: replicate the sampling process from the population at a sample level, i.e., by replacing the population d.f. with a reasonable estimate.

Then, the simplest way to replicate the sampling process at a sampling level simply consists in taking the sample \( s \) (where each unit \( i \) is counted according to its multiplicity) and in performing \( n \) independent, equally probable draws. In practice, a bootstrap sample \( s^* \) is drawn from \( s \) again by srswr of size \( n \). Let \( D_i^* \) represent the multiplicity of unit \( i \) in the bootstrap sample \( s^* \), and let \( D_N^* \) be the \( N \)-variate r.v. with components \( D_i^* \). Then, conditionally on \( D_N \), the r.v. \( D_N^* \) has a multinomial distribution with parameters \( n \) and \( D_i/n, i = 1, \ldots, N \).

As a consequence, if

\[
F_n^*(y) = \frac{1}{n} \sum_{i \in s^*} I_{(-\infty, y]}(Y_i^*) = \frac{1}{n} \sum_{i=1}^N D_i^* I_{(-\infty, y]}(Y_i) \tag{5}
\]

is the bootstrapped e.d.f., then the following two results hold:

\[
E^*[F_n^*(y)|D_N, Y_N] = F_n(y)
\]
\[
V^*[F_n^*(y)|D_N, Y_N] = \frac{1}{n} F_n(y)(1 - F_n(y)).
\]

The main justification of bootstrapping is the asymptotic nature. Consider the empirical processes \( W_N = (\sqrt{N}(F_n(y) - F_{\mathcal{U}_N}(y)); y \in \mathbb{R}) \), \( W_n = (\sqrt{n}(F_n(y) - F_N(y)); y \in \mathbb{R}) \), and the corresponding bootstrapped process \( W_n^* = (\sqrt{N}(F_n^*(y) - F_n(y)); y \in \mathbb{R}) \). As \( N \) increases, the sequence of stochastic processes \( W_N \) converges weakly to a Brownian bridge \( W \) of the scale of \( F_{\mathcal{U}_N} \), namely a Gaussian process with mean function 0 and covariance
Conditionally on $Y_N$, $W_n$ converges weakly to a Brownian bridge $W$ on the scale of $F_{\xi}$ as $N, n$ increase. The same result also holds unconditionally.

E2. $W_N$ weakly converges to a Brownian bridge $W$ on the scale of $F_{\xi}$ as $N$ increases.

E3. $W_n$ and $W_N$ are asymptotically independent.

E4. If $n/N \to f$, with $0 \leq f \leq 1$, then $\sqrt{n}(F_n - F_{\xi})$ converges weakly to $(1 + \sqrt{f})W$, as $n, N$ increase.

E5. Conditionally on $D_N, Y_N, W^*_n$ converges weakly to a Brownian bridge on the scale of $F_{\xi}$ as $N, n$ increase.

The essence of the above results is that the (conditional) distribution of $W^*_n$ asymptotically coincides with the distribution of $W_n$. As a consequence, if we set $\theta^*_n = \theta(F^*_n)$, under the assumption of Hadamard-differentiability of $\theta$ (cfr. [18]), the probability distribution of $\sqrt{n}(\theta_n - \theta_N)$ and that of $\sqrt{n}(\theta(F^*_n) - \theta(F_n))$ converge to the same limit. This is the rationale that explains why the distribution of the estimator $\theta_n$ is approximated by that of $\theta^*_n$.

3. Failure of Efron’s Bootstrap in the Non-i.i.d. Case

Efron’s bootstrap is strictly related to the i.i.d. nature of the random variables (r.v.s) $D_i$s and does not work when the sampling design is without replacement. Consider, for instance, simple random sampling without replacement (srs, for short) design. Suppose that $n/N \to f$, again with $0 \leq f \leq 1$. A “natural” estimator of the population d.f. is still the e.d.f.:

$$F_n(y) = \frac{1}{n} \sum_{i \in s} I_{(-\infty, y]}(Y_i) = \frac{1}{n} \sum_{i=1}^{N} I_{(-\infty, y]}(Y_i),$$

which is, again, an unbiased (and consistent) estimator of both $F_N$ and $F_{\xi}$. Results E1–E4 of Section 2.1 must now be re-formulated in order to take into account the non-independence of r.v.s $I_i$s. More precisely, the following results hold true.

S1. Conditionally on $Y_N$, $W_n$ converges weakly to $\sqrt{1-f}W$, where $W$ is a Brownian bridge on the scale of $F_{\xi}$ as $N, n$ increase. The same result also holds unconditionally.

S2. $W_N$ weakly converges to a Brownian bridge $W$ on the scale of $F_{\xi}$ as $N$ increases.

S3. $W_n$ and $W_N$ are asymptotically independent.

S4. $\sqrt{n}(F_n - F_{\xi})$ converges weakly to $W$, a Brownian bridge on the scale of $F_{\xi}$, as $n, N$ increase.

S5. Conditionally on $D_N$ and $Y_N$, $W^*_n$ converges weakly to a Brownian bridge on the scale of $F_{\xi}$ as $N, n$ increase.

Unless $f = 0$ the asymptotic distribution of $W^*_n$ does not coincide with that of $W_n$. Hence, the probability distribution of $\theta_n$ is generally not well approximated by the distribution of $W^*_n$, neither for finite $n$, nor asymptotically.

Things go even worse for more general sampling designs without replacement, for a simple reason: the e.d.f. is generally an inconsistent estimator of the population d.f. To be concrete, from now on, we focus on sampling designs that are without replacements, of fixed size (i.e., with $I_1 + \cdots + I_N = n$) and with first order inclusion probabilities proportional to $x_i = f(t_{1i}, \ldots, t_{ji})$, $f(\cdot)$ being an appropriate function of the design variables. This covers the important case of $\pi$s sampling designs. In the sequel, the vector of components $x_1, \ldots, x_N$ will be denoted by $X_N$. 

kernel min($F_{\xi}(y_1), F_{\xi}(y_2)) - F_{\xi}(y_1)F_{\xi}(y_2)$. From [16,17], it is easy to see that the following results hold.

E1. Conditionally on $Y_N$, $W_n$ converges weakly to a Brownian bridge $W$ on the scale of $F_{\xi}$ as $N, n$ increase. The same result also holds unconditionally.

E2. $W_N$ weakly converges to a Brownian bridge $W$ on the scale of $F_{\xi}$ as $N$ increases.

E3. $W_n$ and $W_N$ are asymptotically independent.

E4. If $n/N \to f$, with $0 \leq f \leq 1$, then $\sqrt{n}(F_n - F_{\xi})$ converges weakly to $(1 + \sqrt{f})W$, as $n, N$ increase.

E5. Conditionally on $D_N, Y_N, W^*_n$ converges weakly to a Brownian bridge on the scale of $F_{\xi}$ as $N, n$ increase.
In the first place, an elementary computation actually shows that

\[
E_P[F_n(y) \mid Y_N, X_N] = 1/N \sum_{i=1}^n E_P[I_i(X_N) \mid I_{\infty,y}(Y_i)] \\
= 1/N \sum_{i=1}^N 1/\pi_i I_{\infty,y}(Y_i) \\
\neq F_N(y).
\]

As both \( n, N \) increase, the Law of Large Numbers yields

\[
E_P[F_n(y) \mid Y_N, X_N] \rightarrow E_\xi \left[ 1/\pi_i I_{\infty,y}(Y_i) \right] \neq F_\xi(y).
\]

Hence, results E1–E4 do not hold any more, whilst result E5 still holds.

The reason why the original Efron’s i.i.d. bootstrap (sometimes called naive) does not work for general sampling designs is relatively simple. It does not take into account the sampling design according to which the actual sample is drawn. However, we have to stress that this failure is simply due to the i.i.d. nature of the resampling process. The idea on which Efron’s bootstrap rests, namely replicating, at a “sample level” the sampling process from the population is actually correct. What is incorrect is its implementation through simple i.i.d. bootstrap.

As already said in the Introduction, there are several proposals to adapt Efron’s bootstrap to sampling finite populations. In the sequel, we concentrate only on pseudo-population-based bootstrap, essentially for two reasons

1. This is the closest to Efron’s original idea of replicating, at a sample level, the sampling process from the population.
2. This is the only resampling procedure justified by asymptotic arguments similar to those of [17] for Efron’s bootstrap.

4. Accounting for the Sampling Design in Resampling: The Pseudo-Population Approach

Among several techniques that aim at accounting for the sampling design in resampling from finite populations, we consider here the approach based on pseudo-populations. The idea of pseudo-population goes back, at least, to [11] in the case of median estimation essentially under srs when the population size is a multiple of the sample size.

Rather similar ideas are in [12] for srs, again under the condition that the ratio between population size and sample size is a ninteger, and in [13], for stratified random sampling. A major step forward is the paper by [14], where the construction of a pseudo-population is studied under a general \( \pi \)ps sampling design, with general first order inclusion probabilities. In [19], a different approach to the construction of a pseudo-population, very interesting in many respects, is considered.

The pseudo-population approach to resampling can be considered as a two-phase procedure. In the first phase, a pseudo-population (roughly speaking, a prediction of the population) is constructed. In the second phase, a (bootstrap) sample is drawn from the pseudo-population. Broadly speaking, this approach parallels the plug-in principle by Efron.

The pseudo-population is plugged in the sampling process and is used as a “surrogate” of the actual finite population. In the second phase, a sample is drawn from the pseudo-population, according to a sampling design that mimics the original one. In this view, the pseudo-population mimics the real population, and the (re)sampling process from the pseudo-population mimics the (original) sampling process from the real population.
4.1. Pseudo-Populations: Definition

As already said, we confine ourselves to πps sampling designs, with \( \pi_i \propto x_i = f(t_{i1}, \ldots, t_{ij}) \). A pseudo-population is defined as

\[
\{ (N_i^* I_i, y_i, x_i); i = 1, \ldots, N \}
\]

(7)

where \( N_i^* \)'s are integer-valued r.v.s, with (joint) probability distribution \( P_{\text{pred}} \). In practice, Equation (7) means that \( N_i^* I_i \) population units are predicted to have \( y \)-value equal to \( y_i \) and \( x \)-variable \( x_i \), for each sample unit \( i \).

From now on, the familiar bootstrap symbols \( y_i^*, x_i^* \) will be used to denote the \( y \)-value and \( x \)-value of unit \( k \) of the pseudo-population, respectively. Of course \( N_i^* \) units of the pseudo-population satisfy the relationships \( y_k^* = y_i, x_k^* = x_i, i \in s \). The d.f. of the pseudo-population is equal to

\[
F_{N_i^*}(y) = \frac{1}{N^*} \sum_{k=1}^{N^*} I_{(y_k^* \leq y)} = \sum_{i=1}^{N} \frac{N_i^*}{N^*} I_{(y_i \leq y)}, \quad y \in \mathbb{R}
\]

(8)

where

\[
N^* = \sum_{i=1}^{N} N_i^* I_i.
\]

(9)

is the size of the pseudo-population.

An intuitive choice for \( N_i^* \)'s would be \( \pi_i^{-1} \), as remarked, for instance, in [14]. However, such a choice is uneconomical when \( \pi_i^{-1} \) is not an integer. Approaches to the construction of \( N_i^* \) are in [14] and in [19]. General theoretical results, showing that the only correct choice for \( N_i^* \) is to take values that asymptotically behave as \( \pi_i^{-1} \) is in [20]. In that paper, it was essentially shown that expectation (w.r.t. \( P_{\text{pred}} \)) of \( N_i^* \) must be asymptotically equivalent to \( \pi_i^{-1} \):

\[
E[N_i^* | I_N, Y_N, X_N] = \pi_i^{-1} I_i K_{1N}(I_N, Y_N, X_N) \to 1
\]

(10)

as \( N, n \) increase, the symbol \( \to \) in (10) denoting convergence in probability w.r.t. \( I_N \) and for almost all \( y_i \)'s, \( x_i \)'s. Furthermore, in the above mentioned paper additional assumptions on second moments of \( N_i^* \) are made.

A first important example of a pseudo-population satisfying (10) is the Holmberg pseudo-population (cfr. [14]), where:

\[
N_i^* = \lfloor \pi_i^{-1} \rfloor + \epsilon_i
\]

where \( \lfloor x \rfloor \) is the floor function and, conditionally on \( Y_N, X_N, I_N, \epsilon_i \) are independent Bernoulli r.v.s taking value 1 with probability \( r_i = \pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor \) and value 0 with probability \( 1 - r_i \).

A second, important example is the multinomial pseudo-population (cfr. [21]), where, again conditionally on \( I_N \), the joint distribution of \( N_i^* I_i \) is multinomial and corresponds to \( N \) i.i.d. trials, each of them consisting in drawing with replacement a unit from the sample, unit \( i \) having probability \( \pi_i^{-1} I_i / \sum \pi_i^{-1} I_i \) of being selected. Other examples of pseudo-populations, based on various forms of calibration, are in [20].

4.2. Resampling from Pseudo-Populations

Resampling based on pseudo-populations actually parallels Efron’s bootstrap for i.i.d. observations. The basic ideas are relatively simple, once the problem is approached in
terms of an appropriate estimator of the f.p.d.f. To estimate \( F_N \), a simple (but powerful) idea consists in using its Hájek estimator

\[
\hat{F}_H(y) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\pi_i} I_i(-\infty, y](y_i) / \sum_{i=1}^{N} \frac{1}{\pi_i} I_i.
\]  

(11)

As an estimator of a finite population parameter \( \theta_N = \theta(F_N) \), it is then natural to take the statistical functional

\[
\hat{\theta}_H = \theta(\hat{F}_H).
\]  

(12)

A resampling design is a sampling design selecting pseudo-units from the pseudo-population. In the sequel, although it is not strictly necessary, we will assume that the resampling design possesses the same characteristics as the “original” sampling design selecting (real) units from the (real) population. In particular, its first order inclusion probabilities, \( \pi^*_k \), are taken proportional to \( x_k^* \)’s.

Let \( I_k^* \) be the bootstrap sample membership indicator for the pseudo-unit \( k \) of the pseudo-population. The resampled version of \( F_H(y) \) is then equal to

\[
\hat{F}_H^*(y) = \frac{1}{N^*} \sum_{k=1}^{N^*} \frac{1}{\pi_k^*} I_k^*(y_k^*) / \sum_{k=1}^{N^*} \frac{1}{\pi_k^*} I_k^*.
\]  

(13)

On the basis of (13), one may also define the resampled version of \( \hat{\theta}_H \), namely

\[
\hat{\theta}_H^* = \theta(\hat{F}_H^*).
\]

4.3. Resampling Based on Pseudo-Populations: Basics Results for Descriptive Inference

The main theoretical justification for resampling based on pseudo-population is of asymptotic nature, similar, in many respects, to results in [17] for Efron’s bootstrap.

Asymptotics for the distribution of the finite population empirical process \( W_H = (W_H(y); y \in \mathbb{R}) \), where

\[
W_H(y) = \sqrt{n}(\hat{F}_H(y) - F_N(y))
\]

are developed in several papers under different conditions; cfr. [20,22–24]. Here, we confine ourselves to the simplest one, establishing that, under appropriate regularity conditions, as both \( N \) and \( n \) tend to infinity, the following two results hold.

1. Under appropriate regularity conditions, the conditional distribution of \( W_H \), given \( Y_N \) and \( X_N \), converges weakly, as both \( n \) and \( N \) tend to infinity, to a Gaussian process \( W_D \) with null mean function and covariance kernel \( C(y_1, y_2) \). This result, furthermore, holds for a set of sequences of \( y \)’s and \( x \)’s having \( \mathbb{P}_\theta \)-probability 1.

2. If the functional \( \theta(\cdot) \) is Hadamard-differentiable at \( F^*_\xi \) with Hadamard derivative \( \theta^*_F(\cdot) \), then, again conditionally on \( Y_N \) and \( X_N \), \( \sqrt{n}(\hat{\theta}_H - \theta(F_N)) \) tends in distribution to \( \theta^*_F(W_D) \), which is a Normal variate with zero expectation and variance \( \sigma^2_\theta > 0 \).

The rationale behind resampling based on pseudo-population is simple as well as intuitive. The pseudo-population is essentially a “surrogate” of the finite population under consideration, and as both \( N \) and \( n \) increase, their distributions tend to coincide. Hence, at least for a large sample size, the resampling distribution of an estimator should become closer to its actual distribution. This intuition is made rigorous in [20]. Define the resampled empirical process

\[
W_H^* = \sqrt{n}(\hat{F}_H^*-F^*_N).
\]

The following results hold (parallel to results 1 and 2 above).
1*. Under appropriate regularity conditions, the conditional distribution of $W^*_H$, given $Y_N$, $X_N$, $I_N$, converges weakly, as both $n$ and $N$ tend to infinity, to a Gaussian process $W_D$ with a null mean function and covariance kernel $C(y_1, y_2)$. This result, furthermore, holds for a set of sequences of $y_i$s and $t_{ij}$s having $\mathbb{P}_2$-probability 1 and in probability w.r.t. the sampling design.

2*. If the functional $\theta(\cdot)$ is continuously Hadamard-differentiable at $F_\xi$, with Hadamard derivative $\theta'_F(\cdot)$, then, again conditionally on $Y_N$, $X_N$, $I_N$, $\sqrt{n}(\hat{\theta}_H - \theta(F^*_N))$ tends in distribution to $\theta'_F(W_D)$, that turns out to be a Normal variate with zero expectation and variance $\sigma^2 > 0$.

We do not go into detail on the regularity conditions ensuring 1* and 2*. However, it is worth noticing that those results hold true for every pseudo-population satisfying conditions in Section 4.1. With some lack of precision, but more clearly, results 1* and 2* hold for every pseudo-population where $N_i$s asymptotically behave as $\pi^{-1}I_i$s (cfr. relationship (10)).

Even if the conditional (resampling) distribution of $\hat{\theta}_H^*$ is known, its use is not practical for computational reasons. The customary approach essentially consists in resorting to the Law of Large Numbers by making use of independent bootstrap replications. Due to the presence of the finite population, we have now two options.

- **Conditional approach.** A single pseudo-population is constructed, and $M$ independent bootstrap samples are drawn. In this way, $M$ independent replications $\hat{\theta}_{H1}^*, \ldots, \hat{\theta}_{HM}^*$ are generated.

- **Unconditional approach.** $M$ independent pseudo-populations are constructed, and from each of them, a single bootstrap sample is drawn. In this case, $M$ independent replications $\tilde{\theta}_{H1}^*, \ldots, \tilde{\theta}_{HM}^*$ are generated.

As shown in [20], in the case of descriptive inference, conditional and unconditional approaches are asymptotically equivalent. In view of its lower computational burden, a conditional approach seems to be preferable to the unconditional one in descriptive inference.

### 4.4. Resampling Based on Pseudo-Populations: Basics Results for Analytic Inference

The study of a resampling procedure for analytic inference is in principle more complicated than in the case of descriptive inference, essentially because we have to mimic two processes.

- The generation of $y_i$s from the superpopulation model.
- The selection of the sample from the finite population.

In the sequel, as already remarked, we confine ourselves to the simplest case of a superpopulation model where the r.v.s $Y_i$s are i.i.d with common d.f. $F_\xi$. Unlike the case of descriptive inference, where the particular technique according to which the pseudo-population is constructed does not play a relevant role in obtaining asymptotic results, in the present case, the construction of the pseudo-population is relevant. As shown in [25], the only pseudo-population that works for analytical inference is the multinomial one.

Consider now the empirical process

$$\tilde{W}_H = \sqrt{n}(\hat{F}_H - F_\xi)$$

and its resampled version

$$\tilde{W}_H^* = \sqrt{n}(\hat{F}_H^* - \hat{F}_H)$$

The following results (cfr. [25]), which provide a full justification for (multinomial) pseudo-population resampling for analytic inference, hold true.
1. Under appropriate regularity conditions, the (unconditional) distribution of \( \hat{W}_H \)
converges weakly, as both \( n \) and \( N \) tend to infinity to a Gaussian process \( W_A \) with a
null mean function and covariance kernel \( \hat{C}(y_1, y_2) \).

1*. Under appropriate regularity conditions, and conditionally on \( Y_N, X_N, I_N \), the
distribution of \( \hat{W}_H \) converges weakly, as both \( n \) and \( N \) tend to infinity to the same
Gaussian process \( W_A \) with a null mean function and covariance kernel \( \hat{C}(y_1, y_2) \).

2. The limiting process \( W_A \) can be written as \( W_A = W_D + \sqrt{f} W_R \), where \( W_D \) is the
limiting Gaussian process obtained for descriptive inference, \( W_R \) is an independent
Gaussian process (essentially, a Brownian bridge on the scale of \( \mathbb{F}_\xi \)), and \( f \) is the
limiting value of the sampling fraction.

3. If the functional \( \theta(\cdot) \) is Hadamard-differentiable at \( \mathbb{F}_\xi \), with Hadamard derivative
\( \theta'_{\mathbb{F}_\xi}(\cdot) \), then \( \sqrt{n}(\hat{\theta}_H - \theta(\mathbb{F}_\xi)) \) tends in distribution to \( \theta'_{\mathbb{F}_\xi}(W_A) \), that turns out to be a
Normal variate with zero expectation and variance \( \sigma^2_\theta > 0 \).

3*. If the functional \( \theta(\cdot) \) is continuously Hadamard-differentiable at \( \mathbb{F}_\xi \), with Hadamard
derivative \( \theta'_{\mathbb{F}_\xi}(\cdot) \), then, conditionally on \( Y_N, X_N, I_N, W_D, \sqrt{n}(\hat{\theta}_H - \hat{\theta}_I) \) tends in
distribution to the same Normal variate with zero expectation and variance \( \sigma^2_\theta \).

Results 1–3* show that, in analytic inference, there is an extra source of variability,
i.e., \( W_R \), related to the superpopulation model but not depending on the sampling design,
which only affects the term \( W_D \). The smaller the limiting sampling fraction \( f \), the more
negligible the term \( W_R \). As \( f \) tends to zero, results for analytic inference tend to coincide
with the results for descriptive inference.

The above results only hold for multinomial pseudo-populations (with unconditional
approach). The reason is relatively simple: only the multinomial pseudo population (with
unconditional approach) can recover the term \( W_D \) and, hence, the extra variability due to
superpopulation. The problem is negligible when the limiting sampling fraction \( f \) is very
small, but may become relevant for not overly small values of \( f \).

Exactly as in Section 4.3, the use of the exact conditional (resampling) distribution of
\( \hat{\theta}_H^* \) is computationally too difficult. Again, the response consists in generating independent
bootstrap replications. However, in this case, only the unconditional approach works. Hence,
the wide range of options for descriptive inference, in the case of analytic inference
essentially reduces to a single option, namely the multinomial pseudo-population and
unconditional approach.

5. Computational Issues

Use of the pseudo-population approach, despite its many theoretical merits, is
held back by its computational complexity. Real populations could contain millions
of people, and thus the construction of a pseudo-population could be computationally
cumbersome. For this reason, it is of primary interest to develop shortcuts that, while
possessing the fundamental theoretical properties described in the above sections, are
computationally simple to implement because they avoid the physical construction of the
pseudo-population.

The above points are thoroughly discussed in [26], where 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.

An attempt to avoid complications related to integer-valued \( N_i \)'s is in [27], where
non-integer \( N_i \)'s are allowed \( \nu \) the Horvitz–Thompson-based bootstrap (HTB) method.
However, unless the sampling fraction \( n/N \) tends to 0 as \( N \) and \( n \) increase, HTB does not
generally possess the good asymptotic properties outlined in the previous sections.

An interesting computational shortcut is in [28], where the pseudo-population (again
with possibly non-integer \( N_i \)'s) is only implicitly used, and a computational scheme based
on drawings with replacements from the original sample is proposed. Unfortunately,
although the main idea behind that paper is interesting, the proposed bootstrap method
fails to possess good asymptotic properties.
Computational shortcuts, based on ideas similar to those in [28], but based on correct approximations of first order inclusion probabilities, were developed in [29] for descriptive, design-based inference. In particular, in that paper, methodologies based on drawings with replacements from the original sample were proposed, and their merits, from both a theoretical and a computational point of view, were studied.

As remarked by a referee, another drawback of the pseudo-population approach is the apparent necessity to generate and save a large number of bootstrap sample files. However, it is not necessary to save all the bootstrap sample files. Only the original sample file must be saved along with two additional variables for each bootstrap replicate: one variable that contains the number of times each sample unit is used to create the pseudo-population and another one containing the number of times each sample unit has been selected in the bootstrap sample. In other words, it can be implemented similar to methods that rescale the sampling weights.

6. Open Problems and Final Considerations

The pseudo-population approach, despite its merits, requires further development from both the theoretical and computational perspectives. From a theoretical point of view, the results obtained thus far only refer to non-informative single-stage designs. The consideration of multi-stage designs appears as a necessary development as well as the consideration of non-respondent units.

Again, from a theoretical perspective, a major issue is the development of theoretically sound resampling methodologies for informative sampling designs. The major drawback is that, apart from the exception of adaptive designs (cfr. [30]) and the references therein) first order inclusion probabilities can rarely be computed, as these might depend on unobserved quantities. This is what happens, for instance, with most of the network sampling designs that are actually used for hidden populations, where the inclusion probabilities are unknown and depend on unobserved/unknown network links (cfr. [30,31] and the references therein).

From a computational point of view, as indicated earlier, the computational shortcuts developed thus far only work in the case of descriptive inference. The development of theoretically well-founded computational schemes valid for analytic inference is an important issue that deserves further attention.

Author Contributions: Conceptualization, P.L.C. and F.M.; methodology, P.L.C. and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Sapienza (Ateneo) research grant number RM1201729385472F.

Institutional Review Board Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Efron, B. Bootstrap methods: Another look at the jackknife. Ann. Stat. 1979, 7, 1–26. [CrossRef]
2. Mashreghi, Z.; Haziza, D.; Léger, C. A survey of bootstrap methods in finite population sampling. Stat. Surv. 2016, 10, 1–52. [CrossRef]
3. McCarthy, P.J.; Snowden, C.B. The bootstrap and finite population sampling. In Vital and Health Statistics; Public Heath Service Publication, U.S. Government Printing: Washington, DC, USA, 1985; Volume 95, pp. 1–23.
4. Rao, J.N.K.; Wu, C.F.J. Resampling inference with complex survey data. J. Am. Stat. Assoc. 1998, 83, 231–241. [CrossRef]
5. Sitter, R.R. A resampling procedure for complex data. J. Am. Stat. Assoc. 1992, 87, 755–765. [CrossRef]
6. Chatterjee, A. Asymptotic properties of sample quantiles from a finite population. Ann. Inst. Stat. Math. 2011, 63, 157–179. [CrossRef]
7. Rao, J.N.K.; Wu, C.F.J.; Yue, K. Some recent work on resampling methods for complex surveys. Surv. Methodol. 1992, 18, 209–217.
8. Conti, P.L.; Marella, D. Inference for quantiles of a finite population: Asymptotic vs. resampling results. Scand. J. Stat. 2015, 42, 545–561. [CrossRef]
9. Beaumont, J.F.; Patak, Z. On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. Int. Stat. Rev. 2012, 80, 127–148. [CrossRef]
10. Antal, E.; Tillé, Y. A direct bootstrap method for complex sampling designs from a finite population. *J. Am. Stat. Assoc.* 2011, 106, 534–543. [CrossRef]

11. Gross, S.T. Median estimation in sample surveys. In Proceedings of the Section on Survey Research Methods, American Statistical Association, Houston, TX, USA, 11–14 August 1980; pp. 181–184.

12. Chao, M.T.; Lo, S.H. A bootstrap method for finite population. *Sankhya* 1985, 47, 399–405.

13. Booth, J.G.; Butler, R.W.; Hall, P. Bootstrap methods for finite populations. *J. Am. Stat. Assoc.* 1994, 89, 1282–1289. [CrossRef]

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

15. Chauvet, G. Méthodes de Bootstrap en Population Finie. Ph.D. Dissertation, Laboratoire de Statistique d’enquêtes, CREST-ENSAI, Université de Rennes, Rennes, France, 2007.

16. Conti, P.L. On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. *Sankhya B* 2014, 76, 234–259. [CrossRef]

17. Bickel, P.J.; Freedman, D. Some asymptotic theory for the bootstrap. *Ann. Stat.* 1981, 9, 1196–1216. [CrossRef]

18. van der Vaart, A. *Asymptotic Statistics*; Cambridge University Press: Cambridge, UK, 1998.

19. Pfeffermann, D.; Sverchkov, M. Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B* 1999, 61, 166–186.

20. Conti, P.L.; Marella, D.; Mecatti, F.; Andreis, F. A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. *Bernoulli* 2020, 26, 1044–1069. [CrossRef]

21. Pfeffermann, D.; Sverchkov, M. Prediction of finite population totals based on the sample distribution. *Surv. Methodol.* 2004, 30, 79–92.

22. Boistard, H.; Lophuhaä, H.P.; Ruiz-Gazen, A. Functional central limit theorems for single-stage sampling design. *Ann. Stat.* 2017, 45, 1728–1758. [CrossRef]

23. Bertail, P.; Chautru, E.; Clémençon, S. Empirical Processes in Survey Sampling with (Conditional) Poisson Designs. *Scand. J. Stat.* 2017, 44, 97–111. [CrossRef]

24. Han, Q.; Wellner, J.A. Complex sampling designs: Uniform limit theorems and applications. *Ann. Stat.* 2021, 49, 459–485. [CrossRef]

25. Di Iorio, A. Analytic Inference in Finite Population Framework Via Resampling. Unpublished Ph.D. Thesis, Department of Statistical Science, Sapienza Università di Roma, Roma, Italy, 2016.

26. Ranalli, M.G.; Mecatti, F. 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, Alexandria, VA, USA, 2012; pp. 4088–4099.

27. Quatember, A. *Pseudo-Populations—A Basic Concept in Statistical Surveys*; Springer: New York, NY, USA, 2015.

28. Quatember, A. The Finite Population Bootstrap—From the Maximum Likelihood to the Horvitz-Thompson Approach. *Austrian J. Stat.* 2014, 43, 93–102. [CrossRef]

29. Conti, P.L.; Mecatti, F.; Nicolussi, F. Efficient unequal probability resampling from finite populations. *Comput. Stat. Data Anal.* 2022, 167, 107366. [CrossRef]

30. Thompson, S.K. *Sampling*, 3rd ed; Wiley: New York, NY, USA, 2012.

31. Thompson, S.K. Adaptive and Network Sampling for Inference and Interventions in Changing Populations. *J. Surv. Stat. Methodol.* 2017, 5, 1–21. [CrossRef]