A Simplified Efficient and Direct Unequal Probability Resampling

Un semplice Ricampionamento, efficiente e diretto per campioni a probabilitá variabili

Federica Nicolussi, Fulvia Mecatti and Pier Luigi Conti

Abstract In this paper, a new resampling technique for sampling designs with unequal inclusion probabilities is proposed. The basic idea is to use a resampling design based on *ppswor*. Its main properties are studied, and its relationships with other resampling methodologies are discussed.

Abstract *In questo lavoro si introduce una tecnica di ricampionamento valida per disegni campionari con differenti probabilità di inclusione. L'idea di base è di usare un disegno di ricampionamento di tipo ppswor. Le principalei proprietà del metodo sono studiate, e le relazioni con altre metodologie di ricampionamento sono discusse.*

Key words: resampling, finite population, sampling design, ppswor.

1 Background and Contribution

Resampling algorithms are simple and general tools for assessing estimators' accuracy via variance estimation and for producing confidence intervals and p-values. Resampling provides numerical solutions in non-standard challenging inferential setups so that it has a special appeal for dealing with complex sampling designs for finite population. These include the popular without replacement probability proportional to size (πpsWOR) sampling, where every population unit is assigned a specific probability to be included into the final sample, defined as proportional to an available (positive) covariate with the role of auxiliary variable. The Bootstrap,

Federica Nicolussi
Universitá degli Studi di Milano, e-mail: federica.nicolussi@unimi.it

Fulvia Mecatti
Universitá degli Studi di Milano-Bicocca, e-mail: fulvia.mecatti@unimib.it

Pier Luigi Conti
Sapienza Universitá di Roma e-mail: pierluigi.conti@uniroma1.it

likely the most used resampling method originally proposed by Efron [8] for *iid* sample data, does not work in sampling from finite populations, since it cannot deal with the dependence among sample units due to the sampling design. Several modified techniques have been proposed to overcome this problem. In a recent extensive review [12] such proposals are classified into three groups

1. methods based on a *pseudo-population*, where sample units are first used to construct a *replicate* of the parent population and then bootstrap samples are selected into the resulting pseudo-population. Main proposals in this class are [9], [5], [4], [11] and more recently [6] and [7];
2. *direct* bootstrap methods where bootstrap samples are directly selected from the (original) sample or a *re-scaled* version of it. Main proposals in this class are [14], [18] and recently [1];
3. *weighted* bootstrap methods, where a new set of weights is provided to produce bootstrap estimates, by adjusting the (original) design weights. Main contributions of this third type are [15] and [2].

In a recent paper Conti et al. [7] provide a general theory for finite population resampling based on pseudo-population by also proving its asymptotic correctness. The main contribution parallels the asymptotic justification by Bickel and Friedman [3] for the classical *iid* Efron boostrap.

In this paper a new πpsWOR resampling is introduced which is a simplified and computationally more efficient version of the asymptotically correct bootstrap by Conti et al. [7]. The new proposal presents several advantages w.r.t. the large available literature on the topic. First of all, it represents a unified approach to resampling complex samples from finite population. It is in fact a method based on a *pseudo-population*, asymptotically correct according to Conti et al. [7]. However, at the same time it is both a direct bootstrap and a weighted bootstrap, for allowing to select bootstrap samples directly from the original sample on the basis of an appropriate (bootstrap) weighting system. Secondly, it is computationally efficient because it does not require the actual construction of a pseudo-population. In the third place, it is important to notice that the real application of a finite population resampling usually (and certainly for existing methods included in group 1.) involves some sort of rounding or re-scaling, either randomized or systematic, which would affect the entire boostrap performance and ultimately the expected properties of the released bootstrap estimates. The resampling we are proposing does not need any arbitrary rounding and it admits underlying pseudo-population of any size possibly non-integer, along with any real value for the bootstrap weights. A greater precision and possibly efficiency gains are expected as a consequence. Finally, our resampling is very simple to implement, since it requires a unique basic re-sampling design whatever πpsWOR design had generated the available to-be-bootstrapped sample.

2 Notation and Preliminaries

Let $\mathscr{U}_N$ be a finite population of unit $i = 1 \dots N$ from which a sample s is selected under a given design and with pre-fixed size n. Let D_i be the sample membership indicator, i.e. a random variable taking value 1 if $i \in s$ and 0 otherwise, with $n = D_1 + \cdots + D_N$. The (design) expectation $\pi_i = E[D_i] = P(i \in s)$ is the first order inclusion probability. Let $\mathscr{Y}$ be the study variable and $\mathscr{X}$ be an available positive auxiliary variable, with y_i and x_i their value for each population unit, $t_y = \sum_i y_i$ and $t_X = \sum_{i=1}^N x_i$ their population totals. A πpsWOR sampling design is known to be highly efficient whenever $\mathscr{Y}$ is expected to be in a relation of approximate proportionality with $\mathscr{X}$, so that π_i are set to be proportional to the auxiliary variable $\pi_i = nx_i/t_X,\ i = 1 \dots N$. Let $\theta = \theta(F_N)$ be the population quantity to be estimated, where F_N denotes the population distribution function of $\mathscr{Y}$. We focus on the familiar and often used class of estimators that are expressed as functional of an estimator of F_N, namely $\hat{\theta} = \theta\left(\hat{F}\right)$. Such class includes both the popular Horvitz-Thompson and Hájek estimators, respectively given by the following choices to estimate F_N

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

Whatever complex both the sampling design and the analytical structure of θ, a resampling algorithm would produce a (Monte Carlo) estimate of the variance of $\hat{\theta}$ as well as confidence intervals for θ. In Conti et al [7] conditions are given for a resampling algorithm to be asymptotically correct, which implies to be based on a pseudo population and to comprise the following basic steps, where the familiar *star* * notation is adopted to denote bootstrap quantities:

0. Construct a pseudo-population $\mathscr{U}^*_{N^*}$ by replicating (a chosen number) N_i^* of times the values $\{y_i, x_i\}$ associated with every sampled unit $i \in s$. From now on y_k^* and x_k^* will indicate the study and auxiliary values included into the pseudo-population of size N^*, such that N_i^* units $k \in \mathscr{U}^*_{N^*}$ would be *of Type i*, with $i \in s$ and $N^* = \sum_{i \in s} N_i^*$. Finally define the pseudo-population distribution function as

$$F^*_{N^*}(y) = \frac{1}{N^*}\sum_{k=1}^{N^*} I_{(y_k^* \le y)} = \sum_{i=1}^{N} \frac{N_i^*}{N^*} D_i I_{(y_i \le y)},\ \ y \in \mathbb{R} \tag{2}$$

1. Generate M independent bootstrap samples s^* of size n (M chosen sufficiently large) by selecting from $\mathscr{U}^*_{N^*}$ under a (re)sampling design guaranteeing first order inclusion probabilities $\pi_k^* = nx_k^*/t_X^*$ where $t_x^* = \sum_{k=1}^{N^*} x_k^* = \sum_{i \in s}\ N_i^* x_i$.
2. For each bootstrap sample s_m^* compute $\widehat{F}^*_{H,m}$ according to the right term in (1) and thus compute the replicate $\hat{\theta}_m^* = \theta\left(\widehat{F}^*_{H,m}\right),\ m = 1 \dots M$.
3. Compute the M quantities

$$Z^*_{n,m} = \sqrt{n}\left(\widehat{\theta}_m^* - \theta^*\right) = \sqrt{n}\left(\theta(\widehat{F}^*_{H,m}) - \theta(F^*_{N^*})\right) \quad m = 1, \dots, M. \tag{3}$$

Note that (3) provides a bootstrap distribution of $\hat{\theta} = \theta\left(\hat{F}\right)$ - simulated upon M runs - with empirical distribution function given by $\widehat{R}^*_{n,M}(z) = \frac{1}{M}\sum_{m=1}^{M} I_{(Z^*_{n,m} \le z)}$, $z \in \mathbb{R}$ and corresponding pth quantile defined as

$$\widehat{R}^{*-1}_{n,M}(p) = \inf\{z:\ \widehat{R}^*_{n,M}(z) \ge p\},\ \ 0 < p < 1. \tag{4}$$

4. Compute the variance of (3)

$$\widehat{S}^{2*} = \frac{1}{M-1}\sum_{m=1}^{M}\left(Z^*_{n,m} - \overline{Z}^*_M\right)^2 = \frac{n}{M-1}\sum_{m=1}^{M}\left(\widehat{\theta}^*_m - \overline{\theta}^*_M\right)^2 \tag{5}$$

where $\overline{Z}^*_M = \frac{1}{M}\sum_{m=1}^{M} Z^*_{n,m}$ and $\overline{\theta}^*_M = \frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}^*_m$, which is a bootstrap (point) estimate of the variance of estimator $\hat{\theta}$.

6. Finally bootstrap confidence intervals (CI) can be computed; for instance based on percentiles (4)

$$\left[\widehat{\theta} - n^{-1/2}R^{*-1}_{n,M}(1-\alpha/2),\ \widehat{\theta} - n^{-1/2}R^{*-1}_{n,M}(\alpha/2)\right] \tag{6}$$

and based on Standard Normal percentiles and the bootstrap variance estimate at step 3.

$$\left[\widehat{\theta} - n^{-1/2}z_{\alpha/2}\widehat{S}^*,\ \widehat{\theta} + n^{-1/2}z_{\alpha/2}\widehat{S}^*\right] \tag{7}$$

3 The Proposed Method

With the purpose of contributing a new πpsWOR resampling which is a simplified and computationally more efficient version of the asymptotically correct bootstrap recalled above, we aim at compressing the initial steps 0. and 1. in a unique simplified and significantly less time-consuming step. Toward this goal a promising starting point has been a recent Quatember proposal [13]. It focused on a pseudo-population based on the quite natural choice

$$N^*_i = \pi_i^{-1} = t_X/(nx_i) \qquad i \in s \tag{8}$$

which are usually non-integer numbers so that some sort of rounding is needed, the most popular being the Holmberg randomization [11]

$$N^*_{i,Holm} = \lfloor \pi_i^{-1} \rfloor + Bern\left(\pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor\right) \tag{9}$$

The ingenious Quatember solution has two main properties: first of all, the actual construction of the pseudo-population is unnecessary and in fact skipped, so that N^* and every N^*_i are allowed any real number non-necessary integer. Important simplifications realize as a consequence: both the construction of the pseudo-

population and any rounding would be avoided, whether randomized or otherwise. Second the M bootstrap samples s_m^* are selected directly from the (original) sample s by a simple draw-by-draw (with replacement, WR) design, related to the so called drawing-probability-proportional to size (pps). Quatember re-sampling design has initial probability of selecting a unit of Type i, either from the underlying $\mathscr{U}_{N^*}^*$ or directly from s, set equal to

$$p_{i,Q}^* = x_i/t_X^* \tag{10}$$

However it can be proved that Quatember method is not asympotically correct, according to Conti et al. [7]. This is essentially because it fails to address the requirement for the re-sampling inclusion probabilities to be proportional to the auxiliary variable, precisely $\pi_k^* = nx_i/t_X^*$ for all N_i^* units $k \in \mathscr{U}_{N^*}^*$ of Type $i \in s$.

By benefit upon the main idea in Quatember first proposal and by retaining its simplicity and computational efficiency, we now propose some modified versions of $p_{i,Q}^*$ able to address, at least approximatively, such requirement which would lead to asymptotically correct resampling methods beside simplified and efficient. It is important to notice that inclusion probabilities for pps design are not proportional to p_is, and do not have an expression in closed form (see for instance [10], p. 95). We then rely upon three useful approximations, each relating the initial selection probability p_i to the first order inclusion probability π_i. By conditioning on the re-sampling first order inclusion probability to equate nx_i/t_X^*, we derive

$$p_{i,R1}^* \approx \log\left(1-\frac{nx_i}{t_X^*}\right)\Big/\sum_{l\in s} N_l^* \log\left(1-\frac{nx_l}{t_X^*}\right) \tag{11}$$

as a first solution based on [16]. A second solution, computationally heavier, is based on [17] and can be computed via the following iterative algorithm:

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

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

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

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

3. Compute

$$\pi_i^*(m+1) = 1-\exp\{-\xi_n^*(m)p_i^*(m)\},\ \ i \in s \tag{12}$$

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

$$p^*_{i,R2} = p^*_i(m), \;\; i \in s. \tag{13}$$

A third option, based on [10], has led to

$$p^*_{i,H} = \frac{x_i}{t^*_X}\left\{1 + \frac{1}{2}\frac{n-1}{n}\left(\frac{nx_i}{t^*_X} - \overline{\pi}^*\right)\right\} \tag{14}$$

where $\overline{\pi}^* = n^{-1}\sum_{i\in S} N^*_i \pi^{*2}_i$.

Finally we considered a fourth solution based on adjusting the choice (8) rather than the initial selection probability. Notice that eqn. (8) leads to the important property $t^*_X = t_X$, *i.e.* the resulting pseudo-population is calibrated w.r.t. the (real) total of the auxiliary variable. On the other hand, neither (8) nor its randomized version (9) satisfy $\sum_{i\in S} N^*_i = N$, *i.e.* the pseudo-population is not calibrated w.r.t. the population size. Our fourth option is based on fostering a pseudo population calibrated w.r.t. both the (real) population size and (real) total of $\mathscr{X}$. Such double calibration (DCal) is reached by replacing either equation (8) or (9) by

$$N^*_i = \frac{1}{\pi_i} + \frac{(N - \sum \pi_i^{-1})(\sum x_i^2)}{n(\sum x_i^2) - (\sum x_i)^2} - \frac{(N - \sum \pi_i^{-1})(\sum x_i)}{n(\sum x_i^2) - (\sum x_i)^2} x_i \tag{15}$$

that is the exact solution of a quadratic constrained optimization problem, and can be any real number, whether integer or not.

Our new method consists in replacing both steps 0. and 1. of the asymptotically correct resampling algorithm given in Section 2, by the following unique simplified and computationally more efficient step

1. Generate M independent bootstrap samples s^* of size n (M chosen sufficiently large) by selecting from s under a draw-by-draw WR (re)sampling design with conditional probability of selecting a unit of Type i at the jth bootstrap draw given by

$$p_j\left(\text{Type } i | s^*_{j-1}\right) = \frac{\max\{0, (N^*_i - h_{i,j-1})x_i\}}{t^*_X - \sum_{l\in S} h_{l,j-1} x_l} \quad j = 2 \ldots n. \tag{16}$$

   where s^*_{j-1} denotes the bootstrap sub-sample informed by the previous $j-1$ draws, $h_{i,j-1}$ is the number of units of type i selected in the first $j-1$ draw. Notice that, at the first draw (j=1) would hold any of the options illustrated in the present section, either equations (11), (13), (14) or (10) joined with (15).

4 Preliminary Empirical Evidence

A preliminary simulation has been carried out with the purpose of empirically testing the performance of our simplified resampling method according to each of the

4 alternative options illustrated in Section 3, and to compare it with some main competitors available in the literature. The simulated scenarios are composed by two populations of increasing size $N = 200,\ 400$. The study and auxiliary variable $\mathscr{Y}$ and $\mathscr{X}$ were generated according to the same model in [1] leading to circa 80% of correlation. For each scenario, 1000 samples were simulated under a Pareto πps design with 20% sampling fraction. Focusing of the population mean $N^{-1}\ t_y$ as the quantity θ to be estimated, we simulated two familiar estimators: the unbiased Horvitz-Thompson estimator $\hat{\theta}_{HT} = N^{-1}\sum_{i\in s} y_i\pi_i^{-1}$ and the more efficient and asymptotically unbiased Hájek estimator $\hat{\theta}_H = \left(\sum_{i\in s}\pi_i^{-1}\right)^{-1}\ \sum_{i\in s} y_i\pi_i^{-1}$.

For each simulated sample, $M = 1000$ bootstrap runs were performed under 7 different resampling methods: 4 proposed in this paper plus 3 competitors, as described in Table 1. The first competitor, dabbed $Holm$, consists of the asymptotically correct resampling algorithm recalled in Section 2 with the pseudo-population constructed via the Holmberg randomization. It is thus interesting to compare with the resampling proposed here which aim at an equivalent resampling but computationally more efficient for avoiding both the actual construction of the pseudo-population and any rounding. The second competitor, *DirAT* for short, has been recently proposed in the literature as a direct bootstrap neither based on a pseudo-population nor requiring $\pi_i^* \propto x_i$ for every sample unit $i \in s$. Thus, it appears interesting to compare with our methods w.r.t the statistical properties of the bootstrap estimate provided. Finally, the third competitor briefly indicated by Q, is the original proposals by Quatember which has been the starting point for developping our new porposal for a simplified, computationally more efficient and yet asymptotically correct resampling.

Table 1 7 Simulated Resampling Methods

Method	Main features	Reference
Holm	resampling from $\mathscr{U}_{N^*}^*$ with $N_i^* = \lfloor\pi_i^{-1}\rfloor + Bern\left(\pi_i^{-1} - \lfloor\pi_i^{-1}\rfloor\right)$ under the same original sampling design with $\pi_k^* = n x_k^*/t_{X^*}^*$	[11]
DirAT	no $\mathscr{U}_{N^*}^*$, direct resampling into s under a combination of special designs	[1]
Q	direct resampling into s under WR pps design with initial draw probabilities (10) and $N_i^* = \pi_i^{-1}$	[13]
*R*1	direct resampling into s under WR pps design with initial draw probabilities (11) and $N_i^* = \pi_i^{-1}$	new
*R*2	direct resampling into s under WR pps design with initial draw probabilities (13) and $N_i^* = \pi_i^{-1}$	new
H	direct resampling into s under WR pps design with initial draw probabilities (14) and $N_i^* = \pi_i^{-1}$	new
DCal	direct resampling into s under WR pps design with initial draw probabilities (10) and N_i^* as in (15)	new

Expected results from our preliminary simulation are

- a significant/dramatic outperformance of any of the new methods $R1, R2, H$ or $DCal$ over both $Holm$ and $DirAT$ w.r.t. the computational time needed for producing botstrap estimate;
- an essentially equivalent performance of any new methods $R1, R2, H$ or $DCal$ as compared to $Holm$ w.r.t to the properties of the final bootstrap estimates with possibly slight gains due to the possibility to avoid the (randomized) rounding;
- an improvement of all new methods over Q as N and n increases for moderate sampling fraction;
- differences between the performance of the 7 simulated resampling methods able to suggest raccomandations for practical application (beside the computational efficiency).

References

1. Antal E., TilléY.: A direct bootstrap method for complex sampling designs from a finite population. Journal of the American Statistical Association, **206**, 534–543 (2011)
2. Beaumont J- F., Patak Z.: On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Ssampling. International Statistical Review, **80**, 127–148 (2012)
3. Bickel P J., Freedman D.: Some asymptotic theory for the bootstrap. The Annals of Statistics, **9**, 1196–1216 (1981)
4. Booth J G., Butler R W., Hall P.: Bootstrap methods for finite populations. Journal of the American Statistical Association, **89**, 1282–1289 (1994)
5. Chao M -T., Lo S -H.: A bootstrap method for finite population. Sankhya, **47**, 399–405 (1982)
6. Chauvet G.: Méthodes de bootstrap en population finie. Ph.D. Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Universioté de Rennes 2 (2007)
7. Conti P L., Marella D., Mecatti F., Andreis F.: A unified principled framework for resampling based on pseudo-populations: asymptotic theory. Technical Report, Arxiv 1705.03827. Submitted for publication, (2017)
8. Efron B.: Bootstrap methods: another look at the jackknife. The Annals of Statistics, **7**, 1–26 (1979)
9. Gross S T.: Median estimation in sample surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, 181–184 (1980)
10. Hájek J.: Sampling from a finite population, Marcel Dekker, New York (1981)
11. Holmberg A.: A bootstrap approach to probability proportional-to-size sampling. Proceedings of the ASA Section on Survey Research Methods, 378–383 (1998)
12. Mashreghi Z., Haziza D., Léger C.: A survey of bootstrap methods in finite population sampling. Statistics Surveys, **10**, 1–52 (2016)
13. Quatember A.: The Finite Population Bootstrap - from the Maximum Likelihood to the Horvitz-Thompson Approach. Austrian Journal of Statistics, **43**, 93–102 (2014)
14. Rao J N K., Wu C F J.: Resampling inference with complex survey data. Journal of the American Statistical Association, **83**, 231–241 (1988)
15. Rao J N K., Wu C F J., Yue K.: Some recent work on resampling methods for complex surveys. Survey Methodology, **18**, 209–217 (1992)
16. Rosén B.: On sampling with probability proportional to size. Journal of Statistical Planning and Inference, **62**, 159–191 (1997)
17. Rosén B.: On inclusion probabilities for order πps sampling. Journal of Statistical Planning and Inference, **90**, 117–143 (2000)
18. Sitter R P.: A resampling procedure for complex data. Journal of the American Statistical Association, **87**, 755–765 (1992)