Efficient unequal probability resampling from finite populations

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

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

Abstract

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

Keywords: Finite populations, sampling designs, resampling, pseudo-population

1. Introduction

The use of resampling methodologies in sampling from finite populations is of considerable interest. The basic starting point consists in observing that the popular bootstrap technique, originally proposed by [**?**], does not work in sampling from finite populations, because of the dependence among sample units due to the sampling design. Several techniques have been proposed to overcome this problem; a nice, recent review is the paper by [**?**]; cfr. also [**?**], [**?**], [**?**].

Among the resampling techniques for sampling designs with pre-fixed first order inclusion probabilities (πps sampling designs, for short), a special role is played by methodologies based on pseudo-populations; cfr. [**?**] for general aspects, and [**?**] for recent theoretical contributions and a simulation study. A feature which is common to several resampling techniques based on pseudo-populations is their computational burden, that could be very large. This motivates the study of resampling methods for πps sampling designs that:

*Corresponding author
URL: pierluigi.conti@uniroma1.it (Pier Luigi Conti)

- share with methods based on pseudo-populations good properties in terms of variance estimation and coverage probability of confidence intervals;

- have a moderate computational burden.

These points are thoroughly discussed in [**?**], 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 . Interesting steps along this path were considered in [**?**]. Unfortunately, as it will be seen in the sequel, Quatember's proposal does not reproduce, neither exactly nor approximately, at least when the sample size increases with the population size, the first order inclusion probabilities of the sampling design under which sample data were collected (the *original* sampling design, henceforth). In fact, it is intuitively evident that a resampling design, as the sample size and the population size become "large", should be closer and closer to the original sampling design. From a more formal point of view, as shown in [**?**], the key property for the asymptotic correctness of a resampling design is that its first order inclusion probabilities should asymptotically coincide with the first order inclusion probabilities of the original sampling design. In the present paper, a new resampling technique, essentially based on sampling with replacement from the original sample is proposed. The basic idea is to use appropriate drawing probabilities in order to reproduce, at least approximately, prefixed first order inclusion probabilities. Its relationships with resampling based on pseudo-populations will be discussed. The relative merits of the proposed resampling technique will be evaluated through a simulation study.

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

2. Preliminary aspects and notation

Let $\mathcal{U}_N$ be a finite population of size N. A sample $\boldsymbol{s}$ is a subset of $\mathcal{U}_N$. For each unit $i \in \mathcal{U}_N$, let D_i be a Bernoulli random variable (r.v.), such that i is (is not) in the sample $\boldsymbol{s}$ whenever $D_i = 1$ ($D_i = 0$), so that $\boldsymbol{s} = \{i \in \mathcal{U}_N : \ D_i = 1\}$. Denote further by $\boldsymbol{D}_N$ the N-dimensional r.v. of components $D_1, \ldots, D_N$. 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 $n_s = D_1 + \cdots + D_N$.

From now on, the character of interest will be denoted by $\mathcal{Y}$, and y_i is its value for unit i of the population. The population total of character $\mathcal{Y}$ is denoted by

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

and the corresponding population mean by

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

The first order inclusion probabilities are frequently chosen to be proportional to an auxiliary variable $\mathcal{X}$. In symbols: $\pi_i \propto x_i$, where x_i is the value of $\mathcal{X}$ for unit i ($i = 1, \ldots, N$). 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.

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

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

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

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

2.1. ppswr *Sampling design*

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

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

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

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

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

2.2. ppswor *Sampling design*

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

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

First order inclusion probabilities for *ppswor* design are not proportional to p_is, and do not have an expression in closed form; see [**?**], p. 95, where this design in termed *successive sampling*. Useful approximations are given in [**?**], [**?**], [**?**].

Approximation R-1 (cfr. [**?**])

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

Approximation R-2 (cfr. [**?**]) Let ξ_n be the (unique) root of the equation (w.r.t. t):

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

Then, the approximate relationship for inclusion probabilities

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

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

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

Approximation H (cfr. [**?**])

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

where

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

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

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

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

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

This principle can be implemented in a conceptually simple way. Let $\boldsymbol{s}^* = (i_1,\ i_2,\ \ldots,\ i_n)$ be an ordered sequence of not-necessarily distinct types in $\boldsymbol{s}$, and let $\boldsymbol{s}^*_j = (i_1,\ i_2,\ \ldots,\ i_j)$, $j = 1,\ \ldots,\ n-1$. Consider next an arbitrary array of $n \times n$ positive numbers

$$p^*_j(i\,;\ i_1,\ \ldots,\ i_{j-1});\ \ i \in \boldsymbol{s},\ j = 1,\ \ldots,\ n. \tag{7}$$

such that

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

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

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

The scheme defined by (8) is completely general. To be concrete, in the sequel we will focus on a special though important case, namely a sequential drawing scheme, similar to *ppswor*. Let $N^*_i \geq 1$, $i \in \boldsymbol{s}$, be the size (non necessarily integer) of type i, and let

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

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

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

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

with drawing probabilities:

$$p_j(\text{Type } i|\boldsymbol{s}^*_{j-1}) \quad = \quad \frac{\max(0, (N^*_i - h_{i,j-1})p^*_i)}{\sum_{l\in\boldsymbol{s}} \max(0,\ (N^*_l - h_{l,j-1})p^*_l)}, \ \ j = 1, \ldots, n. \tag{10}$$

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

$$0 \le h_{i,j} \le N^*_i \ \ \forall i \in \boldsymbol{s}, \quad \sum_{i\in\boldsymbol{s}} h_{i,j} = j \ \ \forall j = 1, \ldots, n$$

hold.

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

$$p_j(\text{Type } i|\boldsymbol{s}^*_{j-1}) \quad = \quad \frac{\max(0, (1 - h_{i,j-1}/K^*)p^*_i)}{\sum_{l\in\boldsymbol{s}} \max(0,\ (1 - h_{l,j-1}/K^*)p^*_l)}, \ \ j = 1, \ldots, n,$$

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

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

which corresponds to drawing types according to a *ppswr* scheme with drawing probabilities p^*_is.

Of course, there are key points to be clarified, namely the choice of the sizes N^*_is and the choice of the initial drawing probabilities p^*_is, which will be addressed in the subsequent Sections.

4. Relationships with pseudo-populations

The scheme of resampling types, introduced in Section 3, has clear connections with the notion of pseudo-population; cfr. [**?**], [**?**], and, for large sample properties, [**?**]. A *pseudo-population* is essentially a prediction, based on sample data, of the actual population. Each unit k of the pseudo-population takes value (x^*_k, y^*_k) equal to one of the (x_i, y_i) sample pairs. Furthermore, exactly N^*_i units of the pseudo-population take the same values (x_i, y_i), with $i \in \boldsymbol{s}$; equivalently, unit i of the sample is replicated N^*_i times in the pseudo-population. More formally, a pseudo-population is represented as the set $U^* = \{(x_i, y_i, N^*_i);\ i \in \boldsymbol{s}\}$. A unit k of the pseudo-population such that $x^*_k = x_i$ and $y^*_k = x_i$ is said to be of *type* i.

If the size N^*_i of a type i, as introduced in Section 3, is integer, then it is equivalent to a pseudo-population where each sample unit i is replicated N^*_i times. This remark opens the road to different criteria for choosing N^*_is. (cfr [**?**]).

For πps design a popular choice is the [**?**] size.

Holmberg size

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

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

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

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

with

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

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

Horvitz-Thompson size

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

$$N_i^* = \frac{1}{\pi_i} = \frac{t_X}{nx_i}, \;\; i \in \boldsymbol{s}.$$

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

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

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

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

5. Drawing probabilities for resampling

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

population is

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

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

1. Approximation R-1

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

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

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

2. Approximation R-2

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

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

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

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

1. Compute

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

   Go to Step 2.

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

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

Go to Step 3.

3. Compute

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

Go to Step 4.

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

5. Stop. Set

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

2. Approximation H

Taking into account (11), it is seen that

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

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

$$p_{i,H}^* = \frac{x_i}{t_X^*} \left\{ 1 + \frac{1}{2} \frac{n-1}{n} \left(\frac{n x_i}{t_X^*} - \overline{\pi}_2^* \right) \right\} \tag{15}$$

for all N_i^* units of type i, with

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

6. Some theoretical justifications

The goal of the present section is to provide a few theoretical justifications to the resampling scheme developed so far. As shown in [**?**], if the sampling design possesses asymptotically maximal entropy, and if N_i^*s satisfy appropriate regularity conditions (the most important one being that their expectations are asymptotically equivalent to π_i^{-1}s), then the resampling design based on (normalized) Conditional Poisson design, also known as Maximum Entropy design or rejective sampling, is fully justified from and asymptotic viewpoint. As a consequence, it is also justified on the basis of Rao's "scaling problem" already mentioned in Section 3, namely the principle of matching the first two moments of linear statistics. In the sequel, the first and second order inclusion probabilities for the normalized Conditional Poisson

design will be denoted by $\pi^{*R}_{(i)}$, $\pi^{*R}_{(ij)}$ for all pairs of distinct units in the pseudo poputation U^*, of type i and type j, respectively, with $j \neq i$. Of course, $\pi^{*R}_{(i)}$ is equal to nx_i/t^*_X for all units of type i.

Resampling design based on *ppswor* does not possess the same asymptotic justification, although it possess good asymptotic properties: cfr. [**?**]. Moreover it possess good properties with regard to the Rao's principle above, as it will be now illustrated.

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

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

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

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

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

Define first

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

As a consequence of (1.9), (1.10) in [**?**], and taking into account that, up to a term asymptotically negligible,

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

we have

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

and, in view of (18),

$$\Delta^{*S}_{ij} \sim \frac{\pi^{*S}_i(1-\pi^{*S}_i)\pi^{*S}_j(1-\pi^{*S}_j)}{\sum_{i\in\boldsymbol{s}} N^*_i \pi^{*S}_i(1-\pi^{*S}_i)} \left\{1 - \left(1 - \frac{\overline{\pi}^{*S} p^*_i}{\overline{p}^* \pi^{*S}_i}\right)\left(1 - \frac{\overline{\pi}^{*S} p^*_j}{\overline{p}^* \pi^{*S}_j}\right)\right\}$$

$$= \quad \Delta_{ij}^{*R} \left\{ 1 - \left(1 - \frac{\overline{\pi}^* p_i^*}{\overline{p}^* \pi_i^{*R}} \right) \left(1 - \frac{\overline{\pi}^* p_j^*}{\overline{p}^* \pi_j^{*R}} \right) \right\} \tag{20}$$

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

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

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

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

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

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

7. Simulation study

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

1) the Quatember's algorithm [**?**] that is comparable both in terms of being based on a pseudo-

population and, at the same time, being simplified by resampling directly from the (original) sample under a *ppswor* design, as mentioned in Section 1;

2) the Holmberg's method [**?**], that involves a resampling based on a pseudo-population under a randomized version of the Horvitz-Thompson size (see Section 4). However the Holmberg's method requires the actual construction of the pseudo-population and then to resample in it by mimicking the original sampling design.

We explored scenarios composed by six populations of increasing size N from 200 to 5000, with the study variable $\mathcal{Y}$ and the auxiliary variable $\mathcal{X}$ generated according to the model $y_i = \left(12.5 + 3x_i^{1.2} + \sigma\epsilon_i\right)^2 + 4000$ where $x_i \sim |N(0,7)|$, $\epsilon_i \sim N(0,1)$ and $\sigma = 15$, leading to a correlation coefficient approximately equal to 0.8. Selection probabilities are taken proportional to values z_i generated from $Z = \mathcal{Y}^{0.2} \cdot LogN(0, 0.025)$ where $LogN$ denotes a Lognormal probability distribution. These choices match similar simulation works available in recent literature: [**?**], [**?**]. For each population, 1000 samples are simulated under a Pareto sampling design. This latter choice has two prominent reasons: Pareto sampling is practical for being very simple to implement and computationally not demanding and, at the same time, it holds good properties for being high entropy and heuristically recognized to be almost equivalent to the asymptotically maximum entropy Rao-Sampford design ([**?**]). Two sampling fractions n/N have been employed, namely 0.04 and 0.20, with the twofold aim of evaluating small to large finite sample sizes and to enhance the simulation of the Hájek asymptotic setup (see [**?**], Ch. 3).

We investigated the estimation of three population parameters, namely

- *Population mean* $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$;
- *Population median* $Me_Y = Me_Y = \inf\{y : F_N(y) \geq 0.5\}$;
- *Population third quartile* $Q3_Y = \inf\{y : F_N(y) \geq 0.55\}$

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

- $\hat{\bar{Y}}_H = \sum_{i=1}^{N} D_i\pi_i^{-1}y_i / \sum_{i=1}^{N} D_i\pi_i^{-1}$;
- $\hat{Me}_Y = \inf\{y : \hat{F}_H(y) \geq 0.5\}$;
- $\hat{Q3}_Y = \inf\{y : \hat{F}_H(y) \geq 0.75\}$ $y \in \mathbb{R}$

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

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

which is popular in practice because of its unbiasedness, although it is frequently less efficient than the asymptotically unbiased Hájek estimator. The simulated scenarios, are summarized in Table 1.

Table 1: Simulated scenarios

Scenarios						
Population size N	200	400	800	1200	2400	5000
Sampling fraction n/N	0.04 **0.20**					
Sample size n	8 **40**	16 **80**	32 **160**	48 **240**	96 **480**	200 **1000**

For each simulated sample, 1000 bootstrap runs are performed under the 5 resampling methods mentioned above and dubbed R-1, R-2, H, Q and Holm respectively. The methods are compared in terms of both Empirical Coverage (EC) and Average Length (AL) of resampling-based Confidence Intervals (CI). The two most popular bootstrap methods have been used: 1) the bootstrap-percentile method, i.e. by the direct use of the the quantiles of the bootstrap replicates; and 2) the method based on the standard Normal quantiles coupled with the (point) bootstrap estimate of the standard error, dubbed bootstrap-stdN.

Simulation results are summarised in Figures 1 and 2. Graphs show the level of EC for increasing population sizes; sample sizes are proportionally increasing too as a consequence of the fixed sample fractions (4% upper panel, 20% lower panel). AL is represented by the dimension of each points. The solid horizontal line indicates the (nominal) confidence level 95%.

As a general remark, for small sample fraction and sizes (upper panels) all five resampling methods tend to perform similarly for all estimators and for both types of bootstrap CIs. It is noticeable, though, the superiority of H (left panels) against HT (right panels) for estimating the population mean. This is apparent for the bootstrap-stdN CIs (panel (b)) for which H estimation systematically provides better EC, and it is also shown by the bootstrap-percentile CIs (panel (a)) for which, as population size increases, HT estimation tends to produce excessively conservative CIs and larger ALs.

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

Focusing on H estimation and larger sample size (left-bottom panels), simulation results reveal some general pattern. Our new resampling method is associated with the best results, quite uniformly for the three parameters estimated, the three approximations R-1, R-2 and H and both types of CIs. Such empirical evidence is consistent with the theoretical properties illustrated in Section 6. Our proposed bootstrap algorithm seems to be able to improve upon both simulated competitors. It is somewhat more precise than Holmberg's resampling (Holm), possibly because it does not require the randomization step used in Holm to construct the pseudo-population by replicating each sample unit an integer number of times. Our proposed resampling gives also better results than Quatember's method (Q), which seems

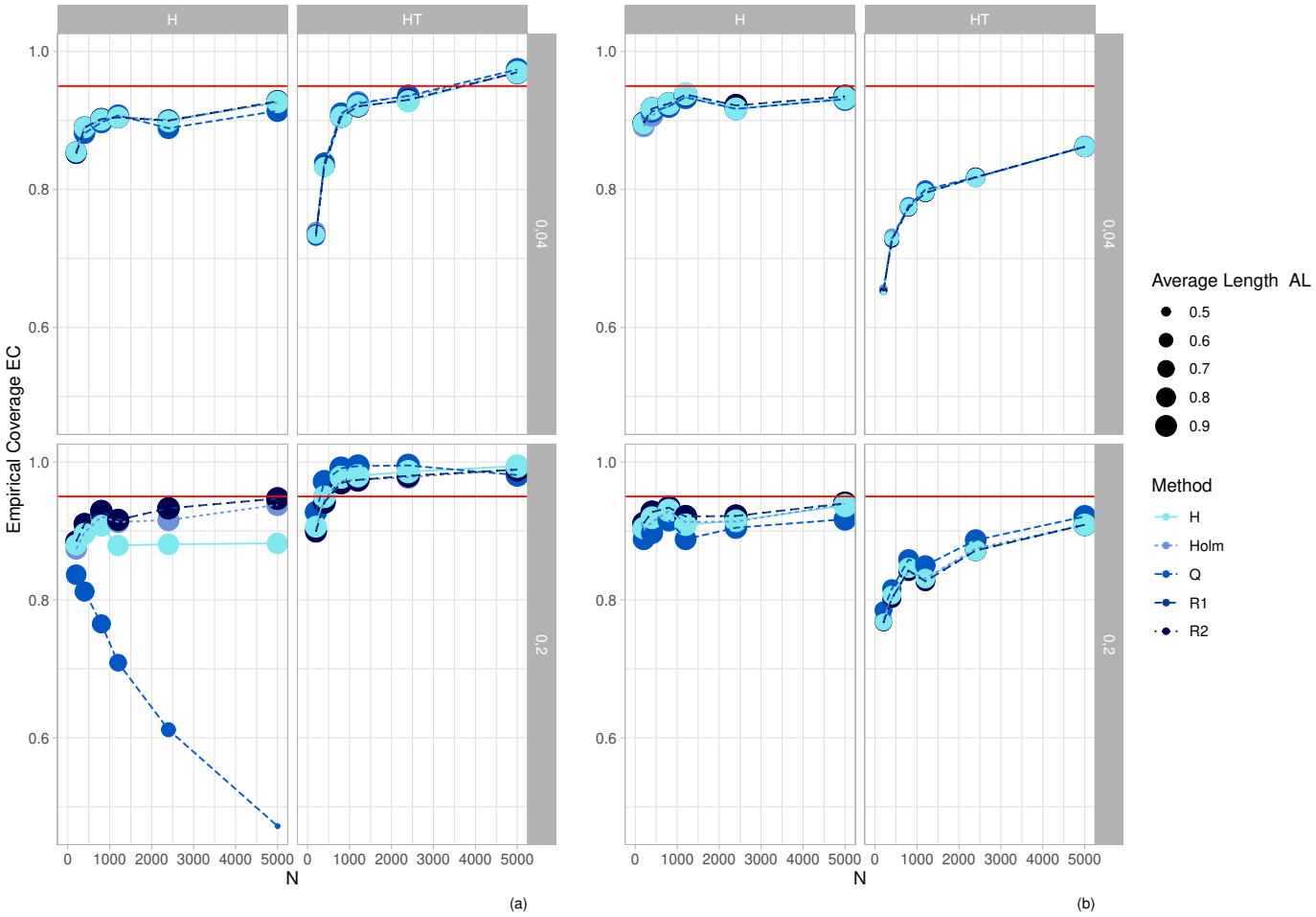

Figure 1: CIs for the population mean for increasing population sizes, bootstrap-percentile (a) and bootstrap-stdN (b), two sample fractions 4% (upper panels) and 20% (lower panels) and two point estimates Hájek (left panels) and Horvitz-Thompson (right panels)

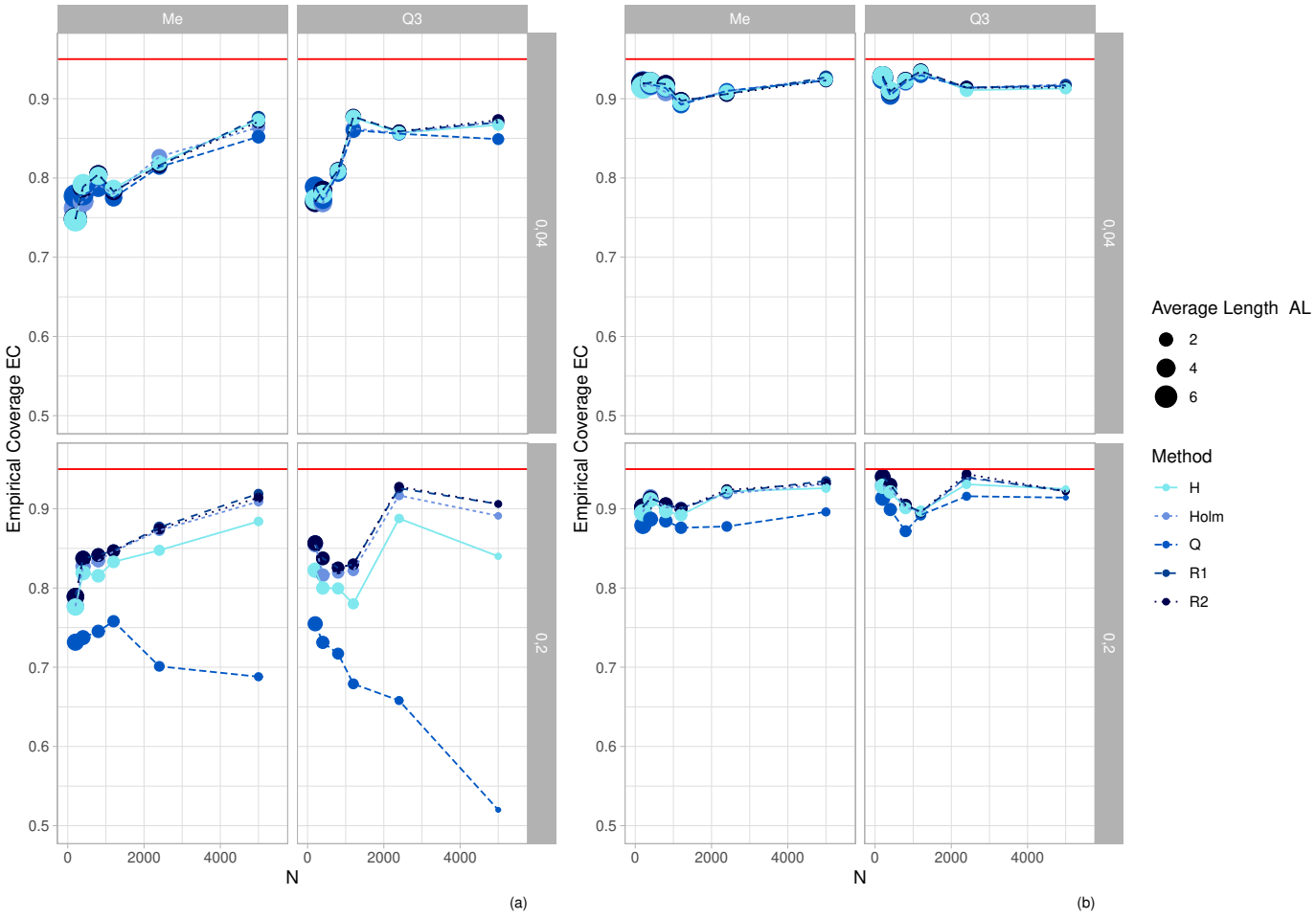

Figure 2: CIs for the population median (Me) (lest panels) and 75% quantile (Q3) (right panels) for increasing population sizes, bootstrap-percentile (a) and bootstrap-stdN (b), two sample fractions 4% (upper panels) and 20% (lower panels)

likely to bear the worst ECs and a tendency to quickly shrinking ALs as N increases. This effect is more enhanced in case of bootstrap-percentile CIs while less evident for bootstrap-stdN CIs. Finally, among the three proposed probability approximations R-1, R-2 and H, the latter shows more erratic and rather weaker performances than both R-1 and R-2 which, at their turn, appear mostly equivalent. In addition, it is worth notice that our proposed resampling method tends to provides good CIs for quantiles, which are population quantity usually tricky to estimate and yet relevant in practice, for instance in studies on national income distribution and social inequalities.

8. Concluding Remarks

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

In order to be implemented, our resampling scheme requires the choice if two quantities, namely (i) the number N_i^* (not necessarily integer) of replicates of each sample unit i, and (ii) the drawing probabilities p_i^*.

As far as the choice of N_i^*s is concerned, the most natural choice appears to be $N_i^* = \pi_i^{-1}$ (HT pseudo-population), that can be implemented even if N_i^*s are not integer. If attention is paid to drawing probabilities, approximations $R-1$ and $R-2$ offer good results, although $R-2$ is slightly heavier from computational viewpoint.

The simulation results of Section 7 also add numerical evidence to the theoretical justifications of Section 6, and explain why our methodology outperform [**?**] original proposal as its main competitor.

Appendix: proofs

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

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

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

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

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

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

$$\begin{aligned}
\frac{\overline{\pi}^*}{\pi_{(i)}^*} &= \frac{\sum_{i \in \boldsymbol{s}} N_i^* \left(f_N^* \frac{x_i}{\overline{X}^*} - f_N^{*2} \frac{x_i^2}{\overline{X}^{*2}}\right)}{f_N^* \frac{x_i}{\overline{X}^*}} \\
&= \frac{N^*}{x_i} \left(\overline{X}^* - f_N^* \frac{m_{2X}^*}{\overline{X}^*}\right).
\end{aligned} \tag{26}$$

Now, as a consequence of (25) and (26), we have then

$$\begin{aligned}
\frac{p_{(i)}^*}{\overline{p}^*} \frac{\overline{\pi}^*}{\pi_{(i)}^*} &= \frac{1 - f_N^* \frac{m_{2X}^*}{\overline{X}^{*2}} + \frac{f_N^*}{2} \frac{x_i}{\overline{X}^*} + O\left(f_N^{*2}\right)}{1 - \frac{f_N^*}{2} \frac{m_{2X}^*}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)} \\
&= \frac{1 + f_N^* \left(\frac{x_i}{2\overline{X}^*} - \frac{m_{2X}^*}{\overline{X}^{*2}}\right) + O\left(f_N^{*2}\right)}{1 - \frac{f_N^*}{2} \frac{m_{2X}^*}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)} \\
&= \left\{1 + f_N^* \left(\frac{x_i}{2\overline{X}^*} - \frac{m_{2X}^*}{\overline{X}^{*2}}\right) + O\left(f_N^{*2}\right)\right\} \left(1 + \frac{f_N^*}{2} \frac{m_{2X}^*}{\overline{X}^{*2}} + O\left(f_N^{*2}\right)\right) \\
&= 1 + \frac{f_N^*}{2} \left(\frac{x_i}{\overline{X}^*} - \frac{m_{2X}^*}{\overline{X}^{*2}}\right) + O\left(f_N^{*2}\right)
\end{aligned} \tag{27}$$

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

$$\begin{aligned}\frac{\Delta_{ij}^{*R}-\Delta_{ij}^{*S}}{\Delta_{ij}^{*R}} &\sim \left(1-\frac{\overline{\pi}^{*}p_{(i)}^{*}}{\overline{p}^{*}\pi_{(i)}^{*R}}\right)\left(1-\frac{\overline{\pi}^{*}p_{(j)}^{*}}{\overline{p}^{*}\pi_{(j)}^{*R}}\right)\\ &= \frac{f_N^{*2}}{4}\left(\frac{x_i}{\overline{X}^{*}}-\frac{m_{2X}^{*}}{\overline{X}^{*2}}\right)\left(\frac{x_j}{\overline{X}^{*}}-\frac{m_{2X}^{*}}{\overline{X}^{*2}}\right)+O\left(f_N^{*3}\right)\end{aligned}$$

from which (23) follows. □

Disclosures and declarations

The authors declare that there is no conflict of interest.

Data transparency

The R code and the data concerning the simulated scenarios are available from the authors on request.

References