Algorithms to Find Exact Inclusion Probabilities for 2Pπps Sampling Designs

Jens Olofsson

July 7, 2010

Abstract

The statistical literature contain several proposals for methods generating fixed size without replacement πps sampling designs. Methods for strict πps designs have rarely been used due to difficulties with implementation. Recently a new method was proposed; the 2Pπps design using a two-phase approach. It was shown that the first-order inclusion probabilities of the 2Pπ design are asymptotically equal to the target inclusion probabilities of a strict πps design.

This paper extends the work on the 2Pπps design and presents algorithms for calculation of exact first- and second-order inclusion probabilities. Starting from a probability mass function (pmf) of the sum of N independent, but not equally distributed Bernoulli variables, the algorithms are based on derived expressions for the pmfs of sums of $N-1$ and $N-2$ variables, respectively.

Exact inclusion probabilities facilitate standard-based inference and provide a tool for studying the properties of the 2Pπps design. Furthermore, empirical results presented show that the properties of the suggested point estimator can be improved using a more general 2Pπps design. In addition, the frequently used Conditional Poisson sampling design is shown to be a special case of this more general 2Pπps design.

1 Introduction

The precision of a sample survey could be increased by using an effective sampling design. One way would be by selecting the population elements into the sample with inclusion probabilities proportional to the variable(s) of interest. However, if that would be possible, there would be no need of a survey since all relevant information would already exist. On the other hand, elements could be selected with probability proportional to some size variable, or function of such, in shape of auxiliary information if such exists, supposed to covary positively with the variable(s) of interest. Such without replacement (WOR) designs with fixed size are called strict πps designs.

Several methods to generate πps designs has been proposed in the statistical literature. See Brewer & Hanif (1983) for an overview and Tillé (2006) for more on both old and more recent πps designs.

In practise strict πps designs have rarely been used, due to difficulties with the implementation. Instead approximative πps designs as the Pareto πps (PAR) and the Conditional Poisson (CPS) have been used, see Rosén (1997*a*,*b*) and Hájek (1964) respectively. However, there exists fast and fairly simple implementations of strict πps designs such as the design presented by Sampford (1967), see Grafström (2009).

Laitila & Olofsson (2010) presented an easily implemented sampling design, the 2Pπps design, based on a two-phase approach yielding a sample of fixed size n proposed to generate a πps sample. It was shown that the first-order inclusion probabilities of the 2Pπps design attain the target inclusion probabilities of a strict πps design asymptotically.

The design was not proposed to replace any of those already existing πps designs. It should rather be seen as an alternative with a simple implementation as its main advantage. A simple implementation of survey designs should not be underestimated since e.g. situations where the frame is not accessible to the statistician him- or herself and the implementation of the design has to be communicated to non- statisticians, are likely to be encountered when working in the field of sample surveys.

In this paper a generalisation of the sampling scheme proposed by Laitila & Olofsson (2010) is presented. When the scheme is used with a Poisson (PO) design as the initial design in first phase, the corresponding sampling design is the 2Pπps design with the Conditional Poisson sampling (CPS) design suggested by Hájek (1964) as a special case.

Algorithms for calculating exact first- and second-order inclusion probabilities of the corresponding design to the presented scheme are derived and presented. This facilitates three different ways of inference; by treating the sample as a true πps sample, by classical two-phase theory or by standard design-based theory, and means for evaluating the proposed design.

2 A Two-phase Sampling Scheme

Neyman (1938) introduced the two-phase (2P), or double (DBL), sampling design as a way of gather information in the first phase necessary for a stratification in the second phase. General formulae for variances and variance estimators, irrespective of sampling designs in each phase, was derived by Särndal & Swensson (1987).

A 2P sampling design can be used in different settings. It can e.g. be used as a way of handling nonresponse; an idea developed by Hansen & Hurwitz

(1946). See also Särndal, Swensson & Wretman (1992, chap. 15), from which the notation here is adopted.

Consider a population $U = \{1, 2, \ldots, N\}$ of N elements and let the value of the variable of interest for element k be denoted by y_k. For sample generation, let n be the predetermined sample size and assume target inclusion probabilities, λ_k, to be proportional to a size variable x_k known for all $k \in U$. The sampling scheme is as follows:

1. Draw a sample, s_0, using a without-replacement (WOR) design with $\Pr(k \in S_0) = \lambda_{ak}$, such that $\sum_{k=1}^{N} \lambda_{ak} = m > 0$.

2. If $n \leq n_{s_0} \leq M$, $M \leq N$, then let $s_a = s_0$ and proceed to step 3. If not, repeat step 1.

3. From the sampled set, s_a, draw a sample s of size n using a simple random sampling WOR (SI) design.

Remark. *Let $\{s_0^i\}_{i=1}^{\infty}$ be an infinite sequence of independent initial samples using any WOR design with $\Pr(k \in S_0^i) = \lambda_{ak}$, such that $\sum_{k=1}^{N} \lambda_{ak} = m > 0$, then the first phase sample $s_a = s_0^\tau$, where $\tau = \min(i : n \leq |s_0^i| \leq M)$.*

Remark. *A sufficient condition for eventually reaching the third step of the scheme is that $\Pr(n \leq |S_0| \leq M) > 0$. Denote this probability with β and let T denote the number of trials before a sample of sufficient size is obtained, then $T \sim Geo(p)$ and $\mathbb{E}(T) = 1/\beta$ and $\mathbb{V}(T) = (1-\beta)/\beta^2$.*

A sample s obtained from a sampling scheme can be interpreted as the outcome of a set-valued random variable S, where its probability function, $\Pr(S = s) = p(s)$, defines the sampling design generated by the sampling scheme. Furthermore, let φ denote the set of all possible samples s such that its cardinality is n, i.e. $\varphi = \{s : |s| = n, s \subseteq U\}$. Given a first phase sample s_a, the probability of selecting a particular subsample s (of size n) in the second phase equals $\binom{|s_a|}{n}^{-1}$. Let $\Omega_s = \{s_a : s \subseteq s_a \subseteq U, |s_a| \leq M\}$. The general sampling design corresponding to the sampling scheme above is then given by

$$p(s) = \sum\nolimits_{s_a \in \Omega_s} p_a(s_a) \binom{|s_a|}{n}^{-1} \tag{1}$$

where $p_a(s_a) = \Pr(S_a = s_a)$ is the sampling design used to generate the first phase sample s_a.

The probability of getting a given sample s using the scheme is dependent, not only on the initial used WOR design but also on the predetermined sample size n and on the constant M. This latter constant should be interpreted as an upper limit of the size of the first phase sample in order to decrease its variation if a non-fixed sampling design is used initially.

The probability of element k belonging to the sample S is given by the sampling design, i.e. $\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s)$, where $s \ni k$ denote the set $\{s : s \in \varphi, k \in s\}$. Although it is possible to calculate the inclusion probabilities by using the probability design it is not always feasible in practise since cardinality of φ will be huge if the population size is large. Fortunately it is possible to derive the expressions analytically.

The first-order inclusion probabilities of the general sampling design corresponding to the scheme above are given by

$$\pi_k = \Pr(k \in S | k \in S_a) \Pr(k \in S_a) \quad k \in U, \tag{2}$$

and the second-order inclusion probabilities by

$$\pi_{kl} = \Pr(k, l \in S | k, l \in S_a) \Pr(k, l \in S_a) \quad \{k, l\} \subset U, k \neq l. \tag{3}$$

3 The 2Pπps design

In Section 2 a general two-phase scheme was proposed. If a PO design with $\lambda_{ak} \propto x_k$ and expected sample size $m = \lfloor \sum_U x_k / \max\{x_k\}_{k=1}^N \rfloor$ is used to draw the initial sample in the first phase and $M = N$, then the corresponding sampling design of the scheme is the 2Pπps sampling design proposed by Laitila & Olofsson (2010). The design was proposed to generate, in an easy way, a sample s of fixed size n with inclusion probabilities proportional to size.

If a PO design is used in the first phase of the sampling scheme presented above the probability function of the corresponding scheme is given by

$$p_{2\mathrm{P}\pi\mathrm{ps}}(s) = c_{2\mathrm{P}\pi\mathrm{ps}} \sum\nolimits_{s_a \in \Omega_s} \prod_{k \in s_a} \lambda_{ak} \prod_{k \in s_a^c} (1 - \lambda_{ak}) \binom{|s_a|}{n}^{-1} \tag{4}$$

where $c_{2\mathrm{P}\pi\mathrm{ps}} = 1/\Pr(n \leq |S_0| \leq M)$, i.e. the reciprocal of the probability of accepting the initial PO sample as a first phase sample.

It is suggested to use the proposed design to, in an easy way, generate a πps sample.

Remark. *Although Laitila & Olofsson (2010) used as expected initial sample size $m = \lfloor \sum_U x_k / \max\{x_k\}_{k=1}^N \rfloor$ and as upper bound of sample size $M = N$ as parameters and here other values, $0 < M \leq N$ and $0 < m \leq N$, are allowed the design given by* (4) *will henceforth be called the 2Pπps sampling design.*

Remark. *It should be noted that if $n = M \leq N$ all the units in the first phase sample are selected with probability one in the second phase of the 2Pπps design. Furthermore, in case that $n = m = M$ the 2Pπps design is identical with the CPS design suggested by Hájek (1964).*

Before continuing, define the membership indicator of the initial sample as

$$I_{ak} = \begin{cases} 1 & \text{if } k \in S_0 \\ 0 & \text{if } k \in S_0^c \end{cases}, \quad k \in U \tag{5}$$

and let $n_{S_0} = \sum_U I_{ak}$.

Lemma 1. *Assume a sampling scheme as defined in Section 2 and let the initial design in the first phase be a PO design with* $\Pr(I_{ak} = 1) = \lambda_{ak}$ *such that* $\sum_U \lambda_{ak} = m$. *Furthermore, let the random event* $\{b \leq |S_0 \setminus D| \leq e\}$ *be denoted by* $C_b^e(D)$, *where* $D \subseteq U$ *or* $D = \emptyset$, *then*

$$\Pr(k \in S_a) = \lambda_{ak} \frac{\Pr(C_{n-1}^{M-1}(k))}{\Pr(C_n^M(\emptyset))}, \quad k \in U \tag{6}$$

and

$$\Pr(k, l \in S_a) = \lambda_{ak}\lambda_{al} \frac{\Pr(C_{n-2}^{M-2}(k, l))}{\Pr(C_n^M(\emptyset))}, \quad \{k, l\} \subset U, k \neq l \tag{7}$$

are the first- and second-order inclusion probabilities, respectively, of the first phase of the 2Pπps design.

Proof. Only (6) will be proved since the proof of (7) is similar. First, let $\{s_0^i\}_{i=1}^{\infty}$ be an infinite sequence of independent samples using a PO design with $\Pr(I_{ak} = 1) = \lambda_{ak}$ and let $s_a = s_0^\tau$, where $\tau = \min(i : n \leq |s_0^i| \leq M)$, then

$$\begin{aligned} \Pr(k \in S_a) &= \Pr(k \in S_0 | C_n^M(\emptyset)) \\ &= \frac{\Pr(I_{ak} = 1, C_n^M(\emptyset))}{\Pr(C_n^M(\emptyset))} \\ &= \frac{\Pr(I_{ak} = 1, C_{n-1}^{M-1}(k))}{\Pr(C_n^M(\emptyset))}, \\ &= \lambda_{ak} \frac{\Pr(C_{n-1}^{M-1}(k))}{\Pr(C_n^M(\emptyset))}. \end{aligned}$$

by the definition of conditional probabilities and the fact that the I_{ak}'s are independent. □

Lemma 1 gives the second factor of π_k and π_{kl}, given by (2) and (3), respectively. In order to derive the first factor remember that the probability of getting a sample s of size n from the realised first phase sample s_a of size n_{s_a} is given by $\binom{|s_a|}{n}^{-1}$. Furthermore, from the sampling scheme proposed it is known that $n \leq |S_a| \leq M$.

Lemma 2. *Let the prerequisites be as in Lemma 1 then*

$$\Pr(k \in S | k \in S_a) = \sum_{i=n}^{M} \frac{n}{i} \frac{\Pr(n_{S_0}^{-k} = i-1)}{\Pr(C_{n-1}^{M-1}(k))}, \quad k \in U \tag{8}$$

where $n_{S_0}^{-k} = |S_0 \setminus \{k\}|$*, and*

$$\Pr(k, l \subset S | k, l \subset S_a) = \sum_{i=n}^{M} \frac{n(n-1)}{i(i-1)} \frac{\Pr(n_{S_0}^{-k,l} = i-2)}{\Pr(C_{n-2}^{M-2}(\{k, l\}))} \tag{9}$$

where $n_{S_0}^{-k,l} = |S_0 \setminus \{k, l\}|$*, are the first- and second-order inclusion probabilities, respectively, in the second phase of the 2Pπps design.*

Proof. Only (8) will be proved, since the proof of (9) is similar. First, let $n_{S_a} = |S_a|$, then

$$\begin{aligned}
\Pr(k \in S | k \in S_a) &= \mathbb{E}_{p_a}\left(\frac{n}{n_{S_a}} | k \in S_a \right) \\
&= \mathbb{E}_{p_a}\left(\frac{n}{n_{S_a}^{-k} + 1} | C_{n-1}^{M-1}(k) \right) \\
&= \sum_{j=n-1}^{M-1} \frac{n}{j+1} \frac{\Pr(n_{S_0}^{-k} = j)}{\Pr(C_{n-1}^{M-1}(k))} \\
&= \sum_{i=n}^{M} \frac{n}{i} \frac{\Pr(n_{S_0}^{-k} = i-1)}{\Pr(C_{n-1}^{M-1}(k))}
\end{aligned}$$

□

Now, it is possible to state the first- and second-order inclusion probabilities jointly over the the two phases of the 2Pπps design.

Theorem 1. *Let the prerequisites be as in Lemma 1 then*

$$\pi_k = \lambda_{ak} \frac{\sum_{i=n}^{M} \frac{n}{i} \Pr(n_{S_0}^{-k} = i-1)}{\Pr(C_n^M(\emptyset))} \tag{10}$$

where $n_{S_0}^{-k} = |S_0 \setminus \{k\}|$*, and*

$$\pi_{kl} = \lambda_{ak} \lambda_{al} \frac{\sum_{i=1}^{M} \frac{n(n-1)}{i(i-1)} \Pr(n_{S_0}^{-k,l} = i-2)}{\Pr(C_n^M(\emptyset))} \tag{11}$$

where $n_{S_0}^{-k,l} = |S_0 \setminus \{k, l\}|$*, are the first- and second-order inclusion probabilities, respectively, of the 2Pπps design.*

Proof. First, let $\{s_0^i\}_{i=1}^{\infty}$ be an infinite sequence of independent initial PO samples and $s_a = s_0^\tau$, where $\tau = \min\{i : n \leq |s_0^i| \leq M\}$, then

$$\begin{aligned}\pi_k &= \Pr(k \in S) \\ &= \Pr(k \in S | k \in S_a) \Pr(k \in S_a) \\ &= \Pr(k \in S | k \in S_a) \Pr(k \in S_0 | C_n^M(\emptyset)) \end{aligned} \tag{12}$$

by the law of total probability and the definition of conditional probability. By applying the first part of Lemma 1 and Lemma 2, respectively, to (12) the result is obtained.

The proof of (11) is similar. □

Remark. *Expressions for the third- and fourth-order inclusion probabilities obtained from the 2Pπps sampling design can be found in Appendix A. The proofs are similar to that of π_k.*

4 Algorithms to find the pmf for the sum of independent Bernoulli distributed random variables

In many situations independent trials with a dichotomous outcome are encountered. Such cases could be seen as Bernoulli trials with a common probability of success, say δ, or trial specific, i.e. δ_k. Define an indicator function

$$I_k = \begin{cases} 1 & \text{if the } k\text{th trial is a success} \\ 0 & \text{otherwise} \end{cases} \qquad k = 1, \ldots, N \tag{13}$$

and let $Z = \sum_{k=1}^{N} I_k$ denote the number of successes out of N trials.

It is well known that $Z \sim Bin(N, \delta)$ if $\Pr(I_k = 1) = \delta$ for all k. If, on the other hand, the probability of success is unequal between trials the probability mass function (pmf) of Z is not possible to write down in a closed-form expression. However, it is possible to compute the pmf recursively, see also e.g. Chen, Dempster & Liu (1994), Aires (1999) and Bondesson, Traat & Lundquist (2006) for similar result.

Lemma 3. *Let $U = \{1, 2, \ldots, N\}$ and $\Omega = \{k : I_k = 1, k \in U\}$. Furthermore, let $\Omega_i = \{\Omega : |\Omega| = i\}$. In addition, let $P_i^{N-|D|}(D)$ denote the probability of the random event $\{|\Omega \setminus D| = i\}$, where $D \subseteq U$ or $D = \emptyset$, then*

$$P_n^N = \sum_{\Omega \in \Omega_n} \prod_{k \in \Omega} \delta_k \prod_{k \in \Omega^c} (1 - \delta_k), \quad n = 0, 1, \ldots, N, \tag{14}$$

which imply, for all $k \in U$,

$$P_n^N = \begin{cases} \prod_{k=1}^N (1-\delta_k) & \text{for } n = 0 \\ \delta_k P_{n-1}^{N-1}(k) + (1-\delta_k) P_n^{N-1}(k) & \text{for } n = 1, 2, \ldots, N-1 \\ \prod_{k=1}^N \delta_k & \text{for } n = N \end{cases}. \quad (15)$$

Proof. If $n = 0$ it implies that neither Bernoulli trial ends with a success, hence $P_0^N = \Pr(\cap_{k=1}^N \{I_k = 0\}) = \prod_{k=1}^N (1-\delta_k)$ since the I_k's are independent $Bin(1, \delta_k)$ distributed random variables. On the other hand, if $n = N$ it implies that $P_N^N = \prod_{k=1}^N \delta_k$ by independence. If $0 < n < N$ it implies, by the law of total probability,

$$\begin{aligned} P_n^N &= \Pr(Z_n^N) \\ &= \Pr(Z_n^N, I_k = 1) + \Pr(Z_n^N, I_k = 0) \\ &= \Pr(Z_{n-1}^{N-1}(k), I_k = 1) + \Pr(Z_n^{N-1}(k), I_k = 0) \\ &= \Pr(Z_{n-1}^{N-1}(k)) \Pr(I_k = 1) + \Pr(Z_n^{N-1}(k)) \Pr(I_k = 0) \\ &= \Pr(Z_{n-1}^{N-1}(k))\delta_k + \Pr(Z_n^{N-1}(k))(1-\delta_k) \end{aligned}$$

where $Z_i^{N-|D|}(D)$ denotes the random event $\{|\Omega \setminus D| = i\}$ and $D \subseteq U$ or $D = \emptyset$. □

The recursion in Lemma 3 is computer intensive in that sense it incorporates all 2^N subsets $\Omega \subseteq U$. In working with Z, where $\Pr(I_k = 1) = \delta_k$, the pmfs such as $\Pr(Z - I_k = n)$, $n = 1, 2, \ldots, N-1$, or $\Pr(Z - (I_k + I_l) = n)$, $n = 1, 2, \ldots, N-2$, could be of interest. It is possible to use Lemma 3 to obtain both sets of such pmfs. However, that would imply a recursion of size 2^{N-1} and 2^{N-2}, respectively. It is proposed to use the results of Lemma 4 and 5 instead, in order to reduce the computational efforts.

Lemma 4. *Let the prerequisites be as in Lemma 3 and let* $\mu_k = 1/(1-\delta_k)$, *then*

$$P_n^{N-1}(k) = \mu_k \sum_{i=0}^{n} (-1)^{n-i} \left(\frac{\delta_k}{1-\delta_k} \right)^{n-i} P_i^N \quad (16)$$

Proof. Proof by induction on n. Let P_n^N be defined as in Lemma 3 and let $n = 0$, which imply that $I_k = 0$ for all $k = 1, 2, \ldots, N$. Then, $P_0^N = P(\cap_{k=1}^N \{I_k = 0\}) = \prod_{k=1}^N (1-\delta_k)$, due to the independence of the I_k's. Now, exclude an arbitrary trial, say trial k. The probability of the remaining $N-1$ Bernoulli trials all fail is given by $P_0^{N-1}(k) = \mu_k P_0^N$, where $\mu_k = 1/(1-\delta_k)$. If $n = 1$, the pmf of Z could be written as

$$P_1^N = \delta_k P_0^{N-1}(k) + (1-\delta_k) P_1^{N-1}(k).$$

By solving for $P_1^{N-1}(k)$ and substituting $P_0^{N-1}(k)$ by $\mu_k P_0^N$ it possible to write

$$\begin{aligned} P_1^{N-1}(k) &= \mu_k(P_1^N - \mu_k\delta_k P_0^N) \\ &= \mu_k \sum_{i=0}^{1}(-1)^{1-i}\left(\frac{\delta_k}{1-\delta_k}\right)^{1-i} P_i^N. \end{aligned}$$

Hence, (16) is true for $n = 1$.

Assume (16) is true for $n = j < N$, then

$$P_j^{N-1}(k) = \mu_k \sum_{i=0}^{j}(-1)^{j-i}\left(\frac{\delta_i}{1-\delta_i}\right)^{j-i} P_i^N.$$

Now, let $n = j + 1 \leq N$, then by Lemma 3

$$P_{j+1}^N = \delta_k P_j^{N-1}(k) + (1-\delta_k)P_{j+1}^{N-1}(k).$$

By solving for $P_{j+1}^{N-1}(k)$ and substituting in the expression above for $P_j^{N-1}(k)$,

$$\begin{aligned} P_{j+1}^{N-1}(k) &= \mu_k(P_{j+1}^N - P_j^{N-1}(k)\delta_k) \\ &= \mu_k\left(P_{j+1}^N - \mu_k\delta_k \sum_{i=0}^{j}(-1)^{j-i}\left(\frac{\delta_i}{1-\delta_i}\right)^{j-i} P_i^N\right) \\ &= \mu_k\left(P_{j+1}^N + (-1)\frac{\delta_k}{1-\delta_k}\sum_{i=0}^{j}(-1)^{j-i}\left(\frac{\delta_i}{1-\delta_i}\right)^{j-i} P_i^N\right) \\ &= \mu_k\left(P_{j+1}^N + \sum_{i=0}^{j}(-1)^{(j+1)-i}\left(\frac{\delta_i}{1-\delta_i}\right)^{(j+1)-i} P_i^N\right) \\ &= \mu_k \sum_{i=0}^{j+1}(-1)^{(j+1)-i}\left(\frac{\delta_i}{1-\delta_i}\right)^{(j+1)-i} P_i^N. \end{aligned}$$

□

Lemma 5. *Let the prerequisites be as in Lemma 3 and let P_n^N be given by Lemma 3 and $P_n^{N-1}(k)$ by Lemma 4, then*

$$\begin{aligned} P_n^{N-2}(k,l) =& \mu_{kl}\sum_{j=0}^{n}(-1)^{n-j}\left(\frac{\delta_l}{1-\delta_l}\right)^{n-j} \\ &\times \sum_{i=0}^{j}(-1)^{j-i}\left(\frac{\delta_k}{1-\delta_k}\right)^{j-i} P_i^N \end{aligned} \tag{17}$$

where $\mu_{kl} = \mu_k\mu_l$.

Proof. Note that $P_n^{N-2}(k,l)$ can be expressed as

$$P_n^{N-2}(k,l) = \mu_l \sum_{j=0}^{n} (-1)^{n-j} \left(\frac{\delta_l}{1-\delta_l} \right)^{n-j} P_j^{N-1}(k)$$

by applying Lemma 4 to the reduced population $U \backslash \{k\}$. By substituting the expression for $P_n^{N-1}(k)$ given by Lemma 4 and letting $\mu_k \mu_l = \mu_{kl}$ the proof is complete. □

By applying the lemmas in this section to the expressions in Theorem 1 it is possible to calculate the first- and second-order inclusion probabilities of the 2Pπps design efficiently.

5 Exact Inclusion Probabilities of the 2Pπps Design

Exact inclusion probabilities of the 2Pπps design facilitates a classical design-based inference, see Särndal et al. (1992). These are possible to compute efficiently by the following theorem.

Theorem 2. *Let the prerequisites be as in Lemma 3, then*

$$\pi_k = \frac{\sum_{i=n}^{M} \frac{n}{i} \sum_{h=0}^{i-1} (-1)^{i-1-h} \alpha_{ak}^{i-h} P_h^N}{\sum_{i=n}^{M} P_i^N} \quad k = 1, 2, \ldots, N, \tag{18}$$

and

$$\pi_{kl} = \frac{\sum_{i=n}^{M} \frac{n(n-1)}{i(i-1)} \sum_{j=0}^{i-2} (-1)^{i-2-j} \alpha_{al}^{i-1-j} \sum_{h=0}^{j} (-1)^{j-h} \alpha_{ak}^{j-h+1} P_h^N}{\sum_{i=n}^{M} P_i^N}, \tag{19}$$

where $\alpha_{ak} = \lambda_{ak}/(1-\lambda_{ak})$.

Proof. By substituting Lemma (4) and Lemma (5) into (10) and (11), respectively and by letting $\alpha_{ak} = \lambda_{ak}/(1-\lambda_{ak})$ and $\alpha_{al} = \lambda_{al}/(1-\lambda_{al})$, respectively, the proof is complete. □

6 Evaluation

In order for a sampling design to be of practical use it should be easily implemented and the parameter(s) of interest should be possible to estimate unbiasedly or nearly so.

Suppose the population total of the variable y is of interest, i.e. $t_y = \sum_U y$. The standard design-based estimator of t_y is given by

$$\hat{t}_{y\pi} = \sum\nolimits_s \frac{y_k}{\pi_k}. \tag{20}$$

Laitila & Olofsson (2010) proposed to use

$$\hat{t}_{y\lambda} = \sum\nolimits_s \frac{y_k}{\lambda_k} \tag{21}$$

as an estimator of t_y when using the 2Pπps design and by regarding the obtained sample as a strict πps sample. Unless $\lambda_k = \pi_k$, using (21) instead of (20) will result in some bias since,

$$\begin{aligned} B(\hat{t}_{y\lambda}) &= \mathbb{E}_p\left(\hat{t}_{y\lambda}\right) - t_y \\ &= \mathbb{E}_p\left(\sum\nolimits_S \frac{y_k}{\lambda_k}\right) - t_y \\ &= \mathbb{E}_p\left(\sum\nolimits_U \frac{I_k y_k}{\lambda_k}\right) - t_y \\ &= \sum\nolimits_U \left(\frac{\mathbb{E}(I_k) y_k}{\lambda_k}\right) - t_y \\ &= \sum\nolimits_U \left(\frac{\pi_k}{\lambda_k} - 1\right) y_k. \end{aligned} \tag{22}$$

With $\pi_k \neq \lambda_k$ it is of interest to know the largest possible bias resulting from using the reciprocal of λ_k instead of π_k as design weights. A conservative measure is given by the upper bound of absolute value of the relative bias of the estimator. If $y > 0$ and $\hat{t}_{y\lambda}$ is used as an estimator of t_y, an upper bound of the absolute value of the relative bias is given by

$$\begin{aligned} |RB(\hat{t}_{y\lambda})| &= \left|\frac{\sum_U y_k b_k}{t_y}\right| \\ &\leq \frac{\sum_U y_k |b_k|}{t_y} \\ &\leq \frac{\max\{|b_k|\}_{k=1}^N \sum_U y_k}{t_y} \\ &= \max\{|b_k|\}_{k=1}^N \\ &= \psi \end{aligned} \tag{23}$$

where $b_k = (\pi_k/\lambda_k - 1)$. See also Rosén (2000) and Aires (2000). Denote this upper bound by ψ. If ψ is small, it indicates that $\pi_k \approx \lambda_k$ for all $k \in U$. In the following the relative bias resulting from using the 2Pπps design with $\hat{t}_{y\lambda}$ as estimator and the corresponding ψ measure are studied using three data sets earlier used in the literature.

Example 1. *In this example one of the vectors in Aires (1999) is used combined with an arbitrary, but given, set of y_k values. The maximum integer-valued sample size possible in order to avoid $\lambda_k > 1$ for at least one $k \in U$ is given by $\lfloor \sum_U x_k / \max\{x_k\}_{k=1}^N \rfloor = 2$. Both λ_k and λ_{ak} are computed using $n = m = 2$. These probabilities does not depend on M. The π_k's are computed when $M = 2$ and $M = 5$, where the first situation corresponds to the CPS design and the latter is the 2Pπps design suggested in Laitila & Olofsson (2010).*

In both situations the first-order inclusion probabilities obtained deviate from the target probabilities corresponding to a strict πps sample for all $k \in U$. This implies that some bias will be expected from using $\hat{t}_{y\lambda}$ given by (21). Given the y-vector, in Table 1, the absolute value of the relative bias is about 2 per cent for both designs. However, for the data at hand, the absolute value of the relative bias is reduced by 22 per cent by using the 2Pπps design compared with the CPS design. Furthermore, using a CPS design $\psi = 0.305$ which is three times as large as a 2Pπps design with $m = 2$ and $M = 5$. See Table 1.

It should also be noted that by using a 2Pπps design compared to a CPS design the expected number of draws until an accepted sample is obtained is reduced as well as its variation. See Table 1.

Table 1: Example 1 population values and probabilities

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{\mathrm{2P\pi ps}}$
1	3	1	0.10000	0.06947	0.10058
2	1	2	0.20000	0.15428	0.20621
3	4	3	0.30000	0.25999	0.31730
4	6	5	0.50000	0.57319	0.55640
5	6	9	0.90000	0.94308	0.81950
		$\sum_U$	2.00000	2.00000	2.00000

Table 2: Example 1 design characteristics and relative bias of $\hat{t}_{y\lambda}$

N	M	m	n	$\lvert\mathrm{RB}(\hat{t}_{y\lambda})\rvert$	ψ	β	$\mathbb{E}(T)$	s.d.(T)
5	2	2	2	0.02563	0.30530	0.43040	2.32342	1.75353
5	5	2	2	0.02097	0.11281	0.70290	1.42268	0.77546

Example 2. *In this example another well known auxiliary vector is used, viz. one of the vectors in Sampford (1967). See Table 2 The maximum integer-valued expected sample size is here equal to 5 and the results are shown in Table 2 – 2.*

When $n = 2$ and x_k is small or high the first-order inclusion probabilities of the 2Pπps design is closer to the target probabilities compared to those of the CPS design If $n = 2$, On the other hand, if x_k is around $\bar{x}_U$ or at its maximum the π_k's obtained from using the CPS design are closer to the target probabilities than those obtained from the 2Pπps design. This pattern becomes more apparent as the sample size increases.

Table 3: Example 2 population values and probabilities, $n = 2$

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	1	2	0.08000	0.07303	0.07411
2	4	2.5	0.10000	0.09243	0.09346
3	2	3.5	0.14000	0.13265	0.13325
4	3	4	0.16000	0.15347	0.15374
5	2	5	0.20000	0.19652	0.19601
6	4	5	0.20000	0.19652	0.19601
7	6	5.5	0.22000	0.21874	0.21785
8	7	6.5	0.26000	0.26446	0.26309
9	6	7	0.28000	0.28791	0.28656
10	10	9	0.36000	0.38427	0.38591
		$\sum_U$	2.00000	2.00000	1.99999

Table 4: Example 2 population values and probabilities, $n = 5$

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	1	2	0.20000	0.17760	0.21472
2	4	2.5	0.25000	0.22735	0.26731
3	2	3.5	0.35000	0.33292	0.37005
4	3	4	0.40000	0.38828	0.41979
5	2	5	0.50000	0.50205	0.51480
6	4	5	0.50000	0.50205	0.51480
7	6	5.5	0.55000	0.55915	0.55960
8	7	6.5	0.65000	0.67046	0.64292
9	6	7	0.70000	0.72375	0.68135
10	10	9	0.90000	0.91639	0.81465
		$\sum_U$	5.00000	5.00000	5.00000

The relative bias, in absolute terms, is less than 1 per cent for both designs, irrespectively of sample size, for given vector of y-values. See Table 2 and Table 2. In general though, from the perspective of ψ, the upper bound of the relative bias, in absolute terms, could be reduced by more than 90 *per cent by using a 2Pπps design compared to the corresponding CPS design when the $n = 5$. See Table 2.*

Table 5: Example 2 design characteristics and the relative bias of $\hat{t}_{y\lambda}$

N	M	m	n	$\lvert\mathrm{RB}(\hat{t}_{y\lambda})\rvert$	ψ	β	$\mathbb{E}(T)$	s.d.(T)
10	2	2	2	0.00461	0.08716	0.30906	3.23560	2.68952
10	10	5	2	0.00481	0.07363	0.62978	1.58787	0.96616
10	5	5	5	0.00157	0.11200	0.27071	3.69402	3.15464
10	10	5	5	0.00641	0.09483	0.63421	1.57676	0.95363

It should also be noted that by using the 2Pπps design the expected number of initial samples before accepting the sample as a first phase sample, as well as its variation, is much smaller than the corresponding parameters of the CPS design.

Example 3. *The data set in the last example is the MU281 population from Särndal et al. (1992). The number of inhabitants in the municipalities in 1975 and 1985 are used as the auxiliary information and as the variable of interest, respectively. For this data set neither n nor m can be larger than 49 in order to avoid $\lambda_k \geq 1$ and $\lambda_{ak} \geq 1$, respectively, for at least one $k \in U$.*

Remark. *In order to shorten the tables the population has been sorted with respect to the values on x_k (P75) and only every 28th observation are presented.*

Table 6: Example 3 population values and probabilities, n=2

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	3	4	0.00117	0.00117	0.00117
29	7	7	0.00205	0.00204	0.00204
57	10	9	0.00264	0.00263	0.00263
85	11	11	0.00323	0.00321	0.00321
113	13	13	0.00381	0.00379	0.00379
141	17	15	0.00440	0.00438	0.00438
169	21	19	0.00557	0.00555	0.00555
197	28	27	0.00792	0.00790	0.00790
225	32	33	0.00968	0.00966	0.00966
253	56	53	0.01555	0.01556	0.01556
281	153	138	0.04048	0.04103	0.04104
		$\sum_U$	2.00000	2.00000	2.00000

In case of $n = 2$, the first-order inclusion probabilities are in practise identical for the two designs. The largest absolute difference is 0.000014 which occurs when x_k attain is maximum value 138. See Table 3. The difference between the first-order inclusion probabilities of the two designs increases as

Table 7: Example 3 population values and probabilities, n = 5

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	3	4	0.00293	0.00291	0.00292
29	7	7	0.00513	0.00510	0.00510
57	10	9	0.00660	0.00656	0.00656
85	11	11	0.00807	0.00802	0.00802
113	13	13	0.00953	0.00949	0.00949
141	17	15	0.01100	0.01095	0.01095
169	21	19	0.01393	0.01388	0.01388
197	28	27	0.01980	0.01974	0.01974
225	32	33	0.02420	0.02415	0.02415
253	56	53	0.03887	0.03891	0.03890
281	153	138	0.10120	0.10252	0.10260
		$\sum_U$	5.00000	5.00000	5.00000

Table 8: Example 3 population values and probabilities, n = 25

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	3	4	0.01467	0.01457	0.01458
29	7	7	0.02567	0.02552	0.02552
57	10	9	0.03300	0.03282	0.03282
85	11	11	0.04033	0.04013	0.04012
113	13	13	0.04767	0.04744	0.04743
141	17	15	0.05500	0.05476	0.05475
169	21	19	0.06967	0.06941	0.06938
197	28	27	0.09900	0.09878	0.09871
225	32	33	0.12100	0.12085	0.12076
253	56	53	0.19434	0.19467	0.19452
281	153	138	0.50601	0.51035	0.51302
		$\sum_U$	25.00000	25.00000	25.00000

the sample size increases. When $n = 49$ the π_k's obtained by using the CPS design is closer to the λ_k's than those obtained by using the 2Pπps design, except when $x_k \in [43, 48]$. However, for the data set at hand, the relative bias is in practise equal to zero for both designs. See Table 3.

The upper bound of the relative bias, given by ψ, is about 1 per cent for both designs, irrespective of sample size, except when $n = 49$. In that case, by using the 2Pπps design, the upper bound is about 7 per cent.

The expected number of initial samples before accepting is again much smaller if a 2Pπps design is used compare to a CPS design, as in the previous two examples.

Table 9: Example 3 population values and probabilities, n = 49

k	y_k	x_k	λ_k	π_k^{CPS}	$\pi_k^{2\mathrm{P}\pi\mathrm{ps}}$
1	3	4	0.02875	0.02857	0.02973
29	7	7	0.05031	0.05003	0.05190
57	10	9	0.06468	0.06436	0.06663
85	11	11	0.07906	0.07870	0.08132
113	13	13	0.09343	0.09305	0.09595
141	17	15	0.10780	0.10741	0.11055
169	21	19	0.13655	0.13618	0.13960
197	28	27	0.19405	0.19383	0.19715
225	32	33	0.23717	0.23715	0.23983
253	56	53	0.38090	0.38194	0.37902
281	153	138	0.99179	0.99197	0.91671
		$\sum_U$	49.00000	49.00000	49.00000

Table 10: Example 3 population parameters and the relative bias of $\hat{t}_{y\lambda}$

N	M	m	n	$\lvert\mathrm{RB}(\hat{t}_{y\lambda})\rvert$	ψ	β	$\mathbb{E}(T)$	s.d.(T)
281	2	2	2	0.000001	0.01351	0.27254	3.66917	3.12948
281	281	49	2	0.000001	0.01385	0.59585	1.67827	1.06692
281	5	5	5	0.000001	0.01297	0.17855	5.60076	5.07620
281	281	49	5	0.000001	0.01385	0.56253	1.77769	1.17579
281	25	25	25	0.000000	0.00857	0.08733	11.45035	10.93893
281	281	49	25	0.000001	0.01384	0.53360	1.87407	1.27988
281	49	49	49	0.000003	0.00625	0.06973	14.34056	13.83153
281	281	49	49	0.000009	0.07570	0.52884	1.89093	1.29795

7 Choice of parameters for the 2Pπps design

Laitila & Olofsson (2010) used $m = \lfloor \sum_U x_k / \max\{x_k\}_{k=1}^N \rfloor$ and $M = N$ as parameters in order for the 2Pπps sampling design to be easy to implement. The same holds true for the result on the 2Pπps design presented in the previous section. However, by the formulation of the sampling scheme in Section 2 it is neither necessary that the expected size of the initial sample in the first phase need to be integer-valued nor $m \geq n$ for the algorithm to work. As long as there is a positive probability of getting an initial sample of sufficient size, i.e. $\Pr(n \leq |S_0| \leq M) > 0$, the algorithm will eventually reach its third step.

Furthermore, it is not necessary for m to be integer-valued as in Laitila & Olofsson (2010) as well as in the examples in the previous section. A different set of parameters for the 2Pπps design, i.e. choice on m and M, for given population and predetermined sample size n, could yield first-

order inclusion probabilities even closer to the target probabilities of a strict πps design. In order to illustrate this a step-wise iteration was done using

Table 11: Choice of parameters, (m, M), for minimising ψ for $\hat{t}_{y\lambda}$, given n, using the MU281 population

n	m	M	$\psi^{2P\pi ps}_{M<N}$	$\psi^{2P\pi ps}_{M=N}$	ψ^{CPS}
1	0.25	5	0.001614	0.013846	0.013678
5	6.25	13	0.000280	0.013846	0.012975
9	11.75	21	0.000323	0.013846	0.012220
13	17.00	34	0.000323	0.013846	0.011409
17	22.00	43	0.000423	0.013846	0.010535
21	26.75	41	0.000543	0.013846	0.009590
25	31.50	46	0.000636	0.013844	0.008568
29	36.25	53	0.000672	0.013794	0.007459
33	40.75	56	0.000820	0.013149	0.006486
37	45.50	72	0.000799	0.009009	0.006443
41	49.25	62	0.001344	0.005181	0.006404
45	49.00	54	0.003560	0.034270	0.006342
49	48.75	49	0.004707	0.075697	0.006253

the same population and variables as in Example 3 in Section 6. Table 7 show for which combination of m and M the theoretical upper bound of the relative bias of the point estimator of the population total in absolute terms is minimised.

The results show that it is possible to obtain a ψ value less than 0.5 percent, which imply that $\pi_k \approx \lambda_k$ for all $k \in U$, for all sample sizes if the expected sample size of the initial PO sample and its upper bound, i.e. m and M respectively, are wisely chosen.

8 Discussion

In this paper a generalisation of the sampling scheme presented by Laitila & Olofsson (2010), based on two-phases, is presented. The design in the first phase of the scheme could be any arbitrary WOR design, whence the design in the second phase is an SI design. A general expression of the corresponding sampling design is given together expressions for the first- and second-order inclusion probabilities.

In the standard setting, when a PO design is used in the first phase, the corresponding design is the 2Pπps design; a design which could be used to generate an approximative strict πps sample in a simple manner which is important since a sampling design sometimes has to be communicated to and/or executed by a non-statistician.

Starting from the pmf of the sum of N independent, but not equally distributed Bernoulli variables, algorithms for calculating the first- and second-order inclusion probabilities of the 2Pπps design were derived based on pmfs of sums of $N-1$ and $N-2$ variables. These algorithms facilitate the 2Pπps to be evaluated as well as dealing with three different ways of inference; by standard design-based theory, by two-phase theory or by treating the sample as a true πps sample.

As shown by the different examples in Section 6 the 2Pπps design yields similar results as the, in practise often used, CPS design. In case of target inclusion probabilities close to one, the CPS design is better than the 2Pπps design in terms of the distance between π_k and λ_k, since, by its formulation, such probabilities are not attainable by the 2Pπps design. This could however be handled in two different ways, depending on the situation at hand. One possible solution is to include those population elements with target inclusion probabilities close to one with probability one into the final sample, i.e. by introducing a take-all stratum. Another possibility is to choose the parameters of the design, m and M, more wisely. However the consequences of either one of these possibilities remains to be further studied.

The formulation of the sampling scheme and corresponding sampling design allows for a choice of a different set of parameters, i.e. m and M, than those used by Laitila & Olofsson (2010) and in the examples. If such a set it wisely chosen it has been shown that it is possible to get an upper bound of the relative bias which in practise equals zero, which in turn imply $\pi_k \approx \lambda_k$. However, how to find such a set in a simple manner without a heavy iteration remains to be solved.

A Third- and fourth-order inclusion probabilities of the 2Pπps sampling design

The third-order inclusion probabilities of the 2Pπps sampling design are given by

$$\begin{aligned}\pi_{klq} &= \Pr(k,l,q \in S) \\ &= \Pr(k,l,q \in S | k,l,q \in S_a)\Pr(k,l,q \in S_a) \\ &= \lambda_{ak}\lambda_{al}\lambda_{aq}\frac{\sum_{n=i}^{M}\binom{n}{i}\binom{n-3}{i-3}^{-1}\Pr(n_{S_0}^{-k,l,q}=i-3)}{\Pr(C_n^M(\emptyset))}\end{aligned} \tag{24}$$

whilst the fourth-order inclusion probabilities of the 2Pπps sampling design are given by

$$\begin{aligned}\pi_{klqr} &= \Pr(k,l,q,r \in S) \\ &= \Pr(k,l,q,r \in S | k,l,q,r \in S_a) \Pr(k,l,q,r \in S_a) \\ &= \lambda_{ak}\lambda_{al}\lambda_{aq}\lambda_{ar} \frac{\sum_{n=i}^{M} \binom{n}{i} \binom{n-4}{i-4}^{-1} \Pr(n_{S_0}^{-k,l,q,r} = i-4)}{\Pr(C_n^M(\emptyset))} \end{aligned} \tag{25}$$

The proofs of the expressions are similar to the proof of the first-order inclusion probabilities given by (10).

References

Aires, N. (1999), 'Algorithms to find exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs', *Methodology and Computing in Applied Probability* **1**(4), 457–469.

Aires, N. (2000), 'Comparions between conditional Poisson sampling and Pareto πps sampling designs', *Journal of Statistical Planning and Inference* **88**, 133–147.

Bondesson, L., Traat, I. & Lundquist, A. (2006), 'Pareto sampling versus Sampford and conditional Poission sampling', *Scandinavian Journal of Statistics* **33**, 699–720.

Brewer, K. & Hanif, M. (1983), *Sampling with Unequal Inclusion Probabilities*, Vol. 15 of *Lecture Notes in Statistics*, Springer-Verlag, New York.

Chen, X. H., Dempster, A. P. & Liu, J. (1994), 'Weighted finite population sampling to maximize entropy', *Biometrika* **81**, 457–469.

Grafström, A. (2009), 'Non-rejective implementations of the Sampford sampling design', *Journal of Statistical Planing and Inference* **139**(6), 2111–2114.

Hájek, J. (1964), 'Asymptotic theory of rejective sampling with varying probabilities from a finite population', *The Annals of Mathematical Statistics* **35**(4), 1491–1523.

Hansen, M. H. & Hurwitz, W. N. (1946), 'The problem of non-response in sample surveys', *Journal of the American Statistical Association* **41**, 517–529.

Laitila, T. & Olofsson, J. (2010), A two-phase sampling scheme and πps designs. manuscript.

Neyman, J. (1938), 'Contribution to the theory of sampling human populations', *Journal of the American Statistical Association* **33**, 101–116.

Rosén, B. (1997*a*), 'Asymptotic theory for order sampling', *Journal of Statistical Planning and Inference* **62**, 135–158.

Rosén, B. (1997*b*), 'On sampling with probability proportional to size', *Journal of Statistical Planning and Inference* **62**, 159–191.

Rosén, B. (2000), 'On inclusion probabilities for order πps sampling', *Journal of Statistical Planning and Inference* **90**, 117–143.

Sampford, M. R. (1967), 'On sampling without replacement with unequal probabilities of selection', *Biometrika* **54**, 499–513.

Särndal, C.-E. & Swensson, B. (1987), 'A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse', *International Statistical Review* **55**, 279–294.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer Series in Statistics, Springer Verlag, New York.

Tillé, Y. (2006), *Sampling Algorithms*, Springer Series in Statistics, Springer.