**Revista Colombiana de Estadística**
Junio 2014, volumen 37, no. 1, pp. 125 a 138

# Algorithms to Calculate Exact Inclusion Probabilities for a Non-Rejective Approximate $\pi$ps Sampling Design

## Algoritmos para calcular probabilidades exactas de inclusión para un diseño de muestreo no rechazable $\pi$pt

Zaizai Yan[a], Yuxia Xue[b]

Science College, Inner Mongolia University of Technology, Hohhot, P. R. China

**Abstract**

AP-design, an efficient non-rejective implementation of the $\pi$ps sampling design, was proposed in the literature as an alternative Poisson sampling scheme. In this paper, we have updated inclusion probabilities formulas in the AP sampling design. The formulas of these inclusion probabilities have been greatly simplified. The proposed results show that the AP design and the algorithms to calculate inclusion probabilities are simple and effective, and the design is possible to be used in practice. Three real examples have also been included to illustrate the performance of these designs.

***Key words*:** AP sampling design, Inclusion probabilities, Poisson sampling.

**Resumen**

Una implementación del diseño de muestreo $\pi$pt, que no es de rechazo, ha sido recientemente propuesta como alternativa al esquema de Poisson. En este trabajo, hemos adaptado las formulas de probabilidades de inclusión en el diseño de muestreo Poisson alternativo (AP por sus siglas en inglés). Estas fórmulas han sido significativamente simplificadas. Los resultados propuestos muestran que el diseño AP y los algoritmos para calcular las probabilidades de inclusión son simples y efectivos, y que el diseño se puede usar en la práctica. Se incluyen tres ejemplos reales para ilustrar el desempeño de la propuesta.

***Palabras clave*:** AP diseño de muestra, probabilidades de inclusión, esquema de Poisson.

[a]Professor. E-mail: zz.yan@163.com
[b]Postgraduate student. E-mail: yuxiaxue_imut@163.com

# 1. Introduction

Unequal probability sampling is frequently used in surveys in order to increase the efficiency in the estimation of the population characteristics. A sampling design without replacement and with unequal inclusion probabilities which are proportional to a size variable, that is known for all units in the population is usually called a $\pi$ps sampling design. The $\pi$ps sampling usually produces more efficient estimates than sampling with equal probabilities. Suppose that the finite population $U$ consists of $N$ units labelled $1, \ldots, N$. An auxiliary variable with value $X_i$ for the unit $i$ is known for all $i = 1, \ldots, N$. Assume that $X_i > 0$, for all $i$ and strict inequality for at least one $i$. It is required to estimate the total $Y = \sum_i Y_i$ where the sum is over $1, \ldots, N$, given a sample of size $n$. Let $p_i = nX_i/X, i = 1, \ldots, N$ be the prescribed inclusion probability parameters with $\sum_{i=1}^{N} p_i = n$ with $X$ its corresponding population total. The problem is how to select a sample with fixed size $n$, so that the probability of each unit $i$ to be included in the sample equals just $p_i$. Many papers have proposed sampling schemes in which the inclusion probability of unit $i$ is $\pi_i$. Some important reference are followings: Sen (1953), Durbin (1967), Brewer (1963), Sampford (1967), Hájek (1964, 1981), Rosén (1997), Aires (1999), Bondesson & Thorburn (2008), Bondesson & Grafström (2011), Grafström (2009), Laitila & Olofsson (2011), Olofsson (2011). Most of the schemes with predetermined inclusion probabilities are either difficult to execute or calculate $\pi_{ij}$, the second order inclusion probability units $i$ and $j$, if $n$ is more than 2. Recently, Zaizai, Miaomiao & Yalu (2013) presented a new approximative $\pi$ps design for fixed sample size $n$ as follows:

1. Draw an initial sample $s_0$, using Poisson sampling design with probabilities $\{p_i\}_1^N$. The size of the initial sample $s_0$ is a random variable denoted by $n_{s_0}$.

2. If $n_{s_0} = n$, then the sampling is finished and the sample $s = s_0$. If $n_{s_0} < n$, then replenish the rest units denoted by $s_1$, its size $n - n_{s_0}$, by simple random sampling without replacement (SRSWOR) design from $U - s_0$, the final sample $s = s_0 \cup s_1$. If $n_{s_0} > n$, then remove $n_{s_0} - n$ units denoted by $s_2$, using the SRSWOR-design, from $s_0$, the final sample $s = s_0 - s_2$. The AP design becomes a non-rejective sampling design. Algorithms for calculating exact first- and second-order inclusion probabilities of the corresponding design are too complex and involve a Jacobi over-relaxation iterative method.

***Note* 1.** We assume that the population is such that $p_i = nX_i/X < 1$, for all $i$. You need to remove the cases where $p_i$ is larger than 1 and then iterative removing further units if necessary

The purpose of this paper is to simplify calculation of the first-order and second-order inclusion probabilities of the AP design. The analytical expressions of inclusion probabilities for the AP design presented in Section 2 are simpler to operate than the original one.

## 2. Inclusion Probabilities of AP Design

Now we discuss inclusion probabilities of the AP sampling design. For convenience, we denote the random variable $\sum_{k\in U,k\neq i} I_k$ as $n_{s_0}^{-i}$, the random variable $\sum_{k\in U,k\neq i,k\neq j} I_k$ as $n_{s_0}^{-ij}$, where $I_k = \begin{cases} 1 & \text{if } k \in s_0 \\ 0 & \text{otherwise} \end{cases}$ for all $k \in U$ are indicators for the Poisson sampling. In order to calculate the first and second-order inclusion probabilities of the AP design, we firstly derive the following Proposition and Lemmas. For convenience, the subset $\{1, 2, \cdots, i\}$ of $U$ is abbreviated as $U_i$ and $Pr(\sum_{\alpha=1}^{i} I_\alpha = j)$ as $P_j^i$ where $j = 0, 1, \ldots, i$; $i = 1, 2, \ldots, N$. Then $Pr(n_{s_0} = \nu) = P_\nu^N$, $\nu = 0, 1, \ldots, N$.

**Proposition 1.** *Keep the same assumptions as above and* $q_i = 1 - p_i$. *Then* $P_0^i = \prod_{\alpha=1}^{i} q_\alpha$; $P_k^i = p_i P_{k-1}^{i-1} + q_i P_k^{i-1}$,$1 \leq k \leq i-1$ *and* $P_i^i = \prod_{\alpha=1}^{i} p_\alpha$.

A proof of proposition 1 can be found in Tillé (2006) and Olofsson (2011).

***Note* 2.** Proposition 1 shows that we can calculate $P_0^i, P_1^i, \ldots, P_i^i$ by using $P_0^{i-1}, P_1^{i-1}, \ldots, P_{i-1}^{i-1}$ with initial values $P_0^1 = q_1$ and $P_1^1 = p_1$. By recursive calculation with respect to $i$, we can finally obtain $P_\nu^N, \nu = 0, 1, \ldots, N$.

**Lemma 1.** *Let* $\mu_k = \frac{1}{1-p_k}$, *then*

$$Pr(n_{s_0}^{-k} = \nu) = \mu_k \sum_{j=0}^{\nu} (-1)^{\nu-j} \Big(\frac{p_k}{1-p_k}\Big)^{\nu-j} P_j^N \tag{1}$$

**Lemma 2.** *Given the assumptions as in Lemma 1, then*

$$Pr(n_{s_0}^{-kl} = \nu) = \mu_k \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \Big(\frac{p_l}{1-p_l}\Big)^{\nu-j} \sum_{t=0}^{j} (-1)^{j-t} \Big(\frac{p_k}{1-p_k}\Big)^{j-t} P_t^N \tag{2}$$

Lemma 1 and Lemma 2 are proved in the appendix. Now we present theorems 1 and 2 which the core results of this paper.

**Theorem 1.** *Under the AP-design, the algorithms for calculating the first-order inclusion probabilities can be written as*

$$\pi_k = \sum_{\nu=0}^{N-1} C_k(\nu) \mu_k \sum_{j=0}^{\nu} (-1)^{\nu-j} \Big(\frac{p_k}{1-p_k}\Big)^{\nu-j} \cdot P_j^N \tag{3}$$

*where* $C_k(\nu) = \begin{cases} \frac{(N-n)p_k + (n-\nu)}{N-\nu} & \nu = 0, \ldots, n-1, \\ \frac{np_k}{\nu+1} & \nu = n, \ldots, N-1 \end{cases}$ *and* $P_j^N = Pr(n_{s_0} = j)$.

**Theorem 2.** *Under the AP-design, the analytical formula of the second-order inclusion probabilities is as follows*

$$\pi_{kl} = \sum_{\nu=0}^{N-2} C_{kl}(\nu) \mu_k \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \Big(\frac{p_l}{1-p_l}\Big)^{\nu-j} \sum_{t=0}^{j} (-1)^{j-t} \Big(\frac{p_k}{1-p_k}\Big)^{j-t} P_t^N \tag{4}$$

*where*

$$C_{kl}(\nu)=\begin{cases} q_k q_l \frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)} + (p_k q_l + p_l q_k)\frac{n-\nu-1}{N-\nu-1} + p_k p_l, & \nu = 0,1,\ldots,n-2, \\ p_k p_l \frac{n(n-1)}{(\nu+2)(\nu+1)}, & \nu = n-1,\ldots,N-2. \end{cases}$$

From Theorem 1 and Theorem 2, we can find that the problem to solve $\pi_k$ and $\pi_{kl}$ may be switched into solving a series of $Pr(n_{s_0} = \nu) = P_\nu^N$, $\nu = 0,1,\ldots,N$. We can recursively calculate $P_\nu^N$ by using Proposition 1. Proofs of Theorems 1 and 2 can be found in the appendix.

# 3. Numerical Examples

The statistical literature contains several proposals for methods generating fixed-size without-replacement $\pi$ps sampling designs. In practice, $\pi$ps designs with sample size $n = 2$ are widely used and fully studied. Due to the difficulties in the implementation and the complexity in computing of inclusion probabilities, application of $\pi$ps designs with sample size $n > 2$ is relatively less. Instead, approximate $\pi$ps designs such as the Conditional Poisson design (CP), two-phase $\pi$ps sampling design (2P$\pi$ps), Rośen (1997)'s Pareto design and Zaizai et al. (2013)'s design (AP) have been used. However, there are fast and fairly simple implementations of strict $\pi$ps designs such as systematic $\pi$ps sampling. Unfortunately, its variance estimation is cumbersome.

## 3.1. A Review of some Sampling Designs

Poisson sampling is a method to generate a sample $s$, which has a random size, from a finite population $U$ consisting of $N$ individuals. Each individual $i$ in the population has a predetermined probability $p_i$ and is included in the sample $s$. A Poisson sample may be obtained by using $N$ independent Bernoulli trials to determine whether the individual under consideration is to be included in the sample $s$ or not. The first-order inclusion probabilities of the individuals are equal to the target inclusion probabilities under the Poisson sampling design. A major drawback with the Poisson design is the randomness of the sample size which has urged statisticians to develop sampling schemes providing fixed size $\pi$ps designs.

Conditional Poisson sampling (CP), also called rejective sampling or maximum entropy sampling, was first introduced by Hájek (1964). It is a fixed size sampling design, without replacement, on a finite population, with unequal inclusion probabilities among the units of the population. It was called rejective sampling because Hájek's implementation amounts to drawing samples with the Poisson sampling design which has a random size until the desired size is chosen. In fact, one can also obtain the conditional Poisson design by drawing samples, with replacement, using a multinomial sampling design and rejecting the samples which hold some units of the population more than one.

Laitila & Olofsson (2011) proposed a new method to generate a sample with fixed size and inclusion probabilities proportional to size, viz. the 2P$\pi$ps design

based on a two-phase approach. Consider a population $U$ of $N$ units. For sample generation, let $n$ be the predetermined sample size and assume target inclusion probabilities, $p_k$, to be proportional to a size variable, $x_k$, known for all $k \in U$. The 2P$\pi$ps sampling scheme is as follows:

1. Draw a sample, $s_0$, using a Poisson design with $p_{ak} \propto x_k$ as inclusion probabilities, with expected sample size $E(n_{s_0}) = \sum_U p_{ak} \geq n$.

2. If the size of $s_0$ is greater than or equal to $n$, then proceed to step 3 and let $s_a = s_0$. If not, repeat step 1.

3. From the sampled set, $s_a$, draw a sample $s$ of size $n$ using an SRSWOR design.

It was shown that the first-order inclusion probabilities of the 2P$\pi$ps design are asymptotically equal to the target inclusion probabilities. But the 2P$\pi$ps design is still a rejective sampling design.

Pareto sampling was introduced by Rośen (1997*a*, 1997*b*). It is a simple method to get a fixed size $\pi$ps sample though with inclusion probabilities only approximately as desired, which can be described as follows: firstly independent random numbers$(U_1, \ldots, U_N)$ from $U(0, 1)$ are generated, one value for each population unit $(i = 1, \ldots, N)$. Then Pareto distributed ranking variables $Q_i = \frac{U_i(1-U_i)}{p_i(1-p_i)}$, where $p_i$ is the targeted inclusion probability for unit $i$ and $\sum p_i = n$, are calculated. Those $n$ units with the smallest Q-values are selected as a $\pi$ps sample with fixed size $n$. Bondesson, Traat & Lundqvist (2006) obtained the formulas of first-order and second-order inclusion probabilities for the Pareto design. The true inclusion probabilities only agree with the target inclusion probabilities approximately.

Zaizai et al. (2013) presented an alternative $\pi$ps design (AP) as Section 1. The AP design is a non-rejective sampling design.

## 3.2. Examples

Since the Horvitz-Thompson estimators under the AP design, CP design and (2P$\pi$ps) design are unbiased, their precision is measured by the variance. However, the ratio estimators mentioned by Kadilar & Cingi (2004) and the traditional ratio estimator are biased, so their precision is measured by mean square error (MSE). In the following section, the estimators and their variances(or MSEs) under the AP design, CP design, 2P$\pi$ps design and SRSWOR are studied using three data sets earlier used in the literature. In this paper the AP design and other designs are applied to three populations in which $y$-values are known, so these variances or MSEs can be calculated exactly. This is only to show the performance of various designs. In practice the $y$-values in an interested population will be unknown, the variance or MSE of an estimator cannot be obtained, but can be estimated from a sample. Then, the precision is measured by estimation of variance or MSE. As far as the Horvitz-Thompson estimators under the AP design, CP design and (2P$\pi$ps) design, the Yates-Grundy variance estimators can be used as the precision. It is unbiased estimator for the true variance.

**Example 1.** We have used the data of Kadilar & Cingi (2004) in this section. However, we have considered the data of only Aegean Region of Turkey, as we are interested in unequal probabilities sampling with fixed sample size here. We have applied our proposed method and other unequal probabilities sampling methods, such as the 2P$\pi$ps sampling design and the CP sampling design on the data of apple production amount (as interest of variate $y$) and number of apple trees (as auxiliary variate $x$) in 105 villages of Aegean Region in 1999 (Source: Institute of Statistics, Republic of Turkey).

For a large size population, we may divide the population into three strata according to size of $X_i$, and the AP-design can be used to get a sample of fixed size within each stratum independently. Let the population be stratified into 3 strata, where sample sizes and population sizes are $(N_1, n_1) = (41, 8)$,$(N_2, n_2) = (41, 8)$ and $(N_3, n_3) = (23, 4)$ respectively. Finally we use stratification sampling technique to build estimation. The relative differences of the inclusion probabilities for the AP-design ,2P$\pi$ps-design and CP-design with respect to target inclusion probabilities can be calculated in each stratum respectively. Then, we can build estimators $\widehat{\overline{Y}}_{HT}^{AP}$ , $\widehat{\overline{Y}}_{HT}^{2P\pi ps}$ and $\widehat{\overline{Y}}_{HT}^{CP}$ of population mean $\overline{Y}$ from Table 1, and the variance of $\widehat{\overline{Y}}_{HT}^{AP}$, $\widehat{\overline{Y}}_{HT}^{2P\pi ps}$ and $\widehat{\overline{Y}}_{HT}^{CP}$ are easily computed, respectively. As mentioned previously, it is of interest to compare the efficiency of using alternative sampling schemes, for example, the 2P$\pi$ps design, AP design, CP design and SRSWOR design. We conclude that the proposed method is more efficient than the 2P$\pi$ps design and SRSWOR design. The empirical comparisons included in Table 1 are of interest. It is noticed that the efficiency of the AP design is almost identical to the 2P$\pi$ps design, but it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004) (Note: The MSEs here are different from the original literature, because the original literature has 106 datum, one of which is a invalid data and is removed, this article has 105 datum). Although the CP design is more efficient than the AP design, the CP design is not easy to implement. The some important advantages of the proposed sampling design are not only its implementation as non-rejective, but also its inclusion probabilities that can be calculated recursively.

Table 1: The variances of the AP design, $2P\pi ps$ design, CP design with $n = 20$, and MSE of SRSWOR ratio estimators in example 1. Aegean Region data.

| Sampling scheme | Method of estimation | Variance (or MSE) |
|---|---|---|
| The AP design | $\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$ | 349150 |
| The 2P$\pi$ps design | $\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$ | 375615 |
| The CP design | $\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$ | 188396 |
| SRSWOR | Upadhyaya-Singh 1 | 2331432 |
| SRSWOR | Upadhyaya-Singh 2 | 2330455 |
| SRSWOR | Singh-Kakran | 2329395 |
| SRSWOR | Sisodia-Dwivedi | 2331304 |
| SRSWOR | Traditional | 2331436 |

***Note* 3.** The AP design still is not an exact πps design. The inclusion probabilities will be larger than intended probabilities for small inclusion probabilities and smaller than intended probabilities for large inclusion probabilities. At the extreme case there will be risks of not selecting units which are intended to be taken with probability 1, and of selecting units with intended inclusion probability 0.

**Example 2.** To analyze the performance of the suggested method in comparison to other methods considered in this paper, a natural population data set from the literature (Singh 1967) is being considered. The descriptions of these populations are given below.

$y$: Percentage of hives affected by disease.

$x$: January average temperature.

We shall consider drawing a sample according to the AP design previously developed. The exact and desired first-order inclusion probabilities are listed in Table 2 and the second-order inclusion probabilities are in Table 3. Then, once we get an AP sample, we can build estimator $\widehat{\overline{Y}}_{HT}^{AP}$ of population mean $\overline{Y}$, and the variance of $\widehat{\overline{Y}}_{HT}^{AP}$ is easily computed.

Table 2: The raw data and the first-order inclusion probabilities for the AP design ,the 2P$\pi ps$ design, the CP design and Pareto design, $N = 10, n = 4$ in example 2. Single data.

| Unit $i$ | $y$ | $x$ | $p$ | $\pi_i^{AP}$ | $\pi_i^{2P\pi ps}$ | $\pi_i^{CP}$ | $\pi_i^{Par}$ |
|---|---|---|---|---|---|---|---|
| 1 | 49 | 35 | 0.3333333 | 0.3445468 | 0.3373678 | 0.3262696 | 0.3327040 |
| 2 | 40 | 35 | 0.3333333 | 0.3445468 | 0.3373678 | 0.3262696 | 0.3327040 |
| 3 | 41 | 38 | 0.3619048 | 0.3682212 | 0.3647676 | 0.3575523 | 0.3614987 |
| 4 | 46 | 40 | 0.3809524 | 0.3840479 | 0.3828163 | 0.3785839 | 0.3807203 |
| 5 | 52 | 40 | 0.3809524 | 0.3840479 | 0.3828163 | 0.3785839 | 0.3807203 |
| 6 | 59 | 42 | 0.4000000 | 0.3999062 | 0.4006775 | 0.3997285 | 0.3999585 |
| 7 | 53 | 44 | 0.4190476 | 0.4157930 | 0.4183399 | 0.4209603 | 0.4192101 |
| 8 | 61 | 46 | 0.4380952 | 0.4317052 | 0.4357925 | 0.4422518 | 0.4384713 |
| 9 | 55 | 50 | 0.4761905 | 0.4635925 | 0.4700272 | 0.4849000 | 0.4770065 |
| 10 | 64 | 50 | 0.4761905 | 0.4635925 | 0.4700272 | 0.4849000 | 0.4770065 |

From Table 4, we see that the proposed method has a smaller variance than the CP design. Although the variance of the 2Pπps design is slightly smaller than proposed method, the AP design is easy to implement and generally applicable. In general, the AP design is extremely efficient and it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004).

**Example 3.** The data we considered here is from 35 Scottish farms in Table 5. Let sample size $n$ be equal to 8. The descriptions of these populations are given below (Asok & Sukhatme 1976, page 916).

$y$: Acreage under oats in 1957.

$x$: Recorded acreage of crops and grass for 1947.

The exact first-order and second-order inclusion probabilities for the AP design, 2Pπps design and CP design are calculated. In this example, the efficiencies for the

TABLE 3: The second-order inclusion probabilities $\pi_{ij}^{AP}$ for the AP design, $N = 10, n = 4$ in example 2. Single data.

| | | | | | Unit $j$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Unit $i$ | | | | | | | | | | |
| 1 | 0.34455 | 0.09537 | 0.10268 | 0.10764 | 0.10764 | 0.11267 | 0.11777 | 0.12293 | 0.13347 | 0.13347 |
| 2 | 0.09537 | 0.34455 | 0.10268 | 0.10764 | 0.10764 | 0.11267 | 0.11777 | 0.12293 | 0.13347 | 0.13347 |
| 3 | 0.10268 | 0.10268 | 0.36822 | 0.11588 | 0.11588 | 0.12128 | 0.12675 | 0.13230 | 0.14361 | 0.14361 |
| 4 | 0.10764 | 0.10764 | 0.11588 | 0.38405 | 0.12146 | 0.12711 | 0.13284 | 0.13864 | 0.15047 | 0.15047 |
| 5 | 0.10764 | 0.10764 | 0.11588 | 0.12146 | 0.38405 | 0.12711 | 0.13284 | 0.13864 | 0.15047 | 0.15047 |
| 6 | 0.11267 | 0.11267 | 0.12128 | 0.12711 | 0.12711 | 0.39991 | 0.13901 | 0.14506 | 0.15740 | 0.15740 |
| 7 | 0.11777 | 0.11777 | 0.12675 | 0.13284 | 0.13284 | 0.13901 | 0.41579 | 0.15156 | 0.16442 | 0.16442 |
| 8 | 0.12293 | 0.12293 | 0.13230 | 0.13864 | 0.13864 | 0.14506 | 0.15156 | 0.43171 | 0.17152 | 0.17152 |
| 9 | 0.13347 | 0.13347 | 0.14361 | 0.15047 | 0.15047 | 0.15740 | 0.16442 | 0.17152 | 0.46359 | 0.18595 |
| 10 | 0.13347 | 0.13347 | 0.14361 | 0.15047 | 0.15047 | 0.15740 | 0.16442 | 0.17152 | 0.18595 | 0.46359 |

TABLE 4: The variances of the AP design, 2P$\pi$ps design, CP design and Pareto design with $n = 4$ and MSE of SRSWOR ratio estimators in example 2. Single data.

| Sampling scheme | Method of estimation | Variance(or MSE) |
|---|---|---|
| The AP design | $\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$ | 3.8268 |
| The 2P$\pi$ps design | $\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$ | 3.7047 |
| The CP design | $\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$ | 3.8681 |
| The Pareto design | $\widehat{\overline{Y}}_{HT}^{Par} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{Par}$ | 3.7334 |
| SRSWOR | Upadhyaya-Singh 1 | 10.5488 |
| SRSWOR | Upadhyaya-Singh 2 | 15.6308 |
| SRSWOR | Singh-Kakran | 10.9737 |
| SRSWOR | Sisodia-Dwivedi | 10.4738 |
| SRSWOR | Traditional | 10.5164 |

AP design, CP design and 2P$\pi$ps design are compared. From the results of Table 6, we conclude that the AP design is more efficient than the CP design. Since the CP design and 2P$\pi$ ps design are far more complex than the AP design, the proposed design is significantly better than the CP design and 2P$\pi$ ps design and it is significantly higher than ratio estimators of the SRSWOR design mentioned by Kadilar & Cingi (2004).

A primary purpose of this paper is to extend the theory of finite sampling with unequal probabilities. Although the study variable $y$ of the data presented in Table 5 is often unknown in the real world, they do indicate that substantial reductions in variance can be obtained by using the AP design (Table 1, 4 6). It is the opinion of the authors that the technique suggested in this paper may be an implemented utility in the real world for unknown study variable $y$. Hence, the proposed method has potential application value.

TABLE 5: Recorded Acreage of Crops and Grass for 1947 and Acreage Under Oats in 1957 for 35 Farms in Orkney in example 3. Scottish forms data.

| Farm No. | $x$ | $y$ | Farm No. | $x$ | $y$ | Farm No. | $x$ | $y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 17 | 13 | 78 | 23 | 25 | 209 | 70 |
| 2 | 50 | 17 | 14 | 90 | 0 | 26 | 240 | 28 |
| 3 | 52 | 10 | 15 | 91 | 27 | 27 | 274 | 62 |
| 4 | 58 | 16 | 16 | 92 | 34 | 28 | 300 | 59 |
| 5 | 60 | 6 | 17 | 96 | 25 | 29 | 303 | 66 |
| 6 | 60 | 15 | 18 | 110 | 24 | 30 | 311 | 58 |
| 7 | 62 | 20 | 19 | 140 | 43 | 31 | 324 | 128 |
| 8 | 65 | 18 | 20 | 140 | 48 | 32 | 330 | 38 |
| 9 | 65 | 14 | 21 | 156 | 44 | 33 | 356 | 69 |
| 10 | 68 | 20 | 22 | 156 | 45 | 34 | 410 | 72 |
| 11 | 71 | 24 | 23 | 190 | 60 | 35 | 430 | 103 |
| 12 | 74 | 18 | 24 | 198 | 63 | | | |

TABLE 6: The variances of the AP design, 2Pπps design, CP design with $n = 8$ and MSE of SRSWOR ratio estimators in example 3. Scottish forms data.

| Sampling scheme | Method of estimation | Variance(or MSE) |
|---|---|---|
| The AP design | $\widehat{\overline{Y}}_{HT}^{AP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{AP}$ | 15.7658 |
| The 2Pπps design | $\widehat{\overline{Y}}_{HT}^{2P\pi ps} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{2P\pi ps}$ | 15.3746 |
| The CP design | $\widehat{\overline{Y}}_{HT}^{CP} = \frac{1}{N}\sum_{i\in s} y_i/\pi_i^{CP}$ | 16.8456 |
| SRSWOR | Upadhyaya-Singh 1 | 99.4516 |
| SRSWOR | Upadhyaya-Singh 2 | 99.5016 |
| SRSWOR | Singh-Kakran | 99.2217 |
| SRSWOR | Sisodia-Dwivedi | 97.9005 |
| SRSWOR | Traditional | 98.5479 |

# 4. Conclusions

We have shown that it is feasible to calculate the first-order and second-order inclusion probabilities in the AP design. Expressions for the third-order and fourth-order inclusion probabilities under the AP sampling design can be obtained. The proofs are similar to that of $\pi_k$.

This study shows that the AP design possesses approximately the same efficiency with the CP design and 2Pπps design. But the AP design is a non-rejective sampling design and very close to the strict πps design. First and second-order inclusion probabilities can be accurately calculated by using the formula given in this paper. From these numerical illustrations, it is deduced that there is considerable gain in efficiency by using the Horvitz-Thompson estimator under the AP design over the other ratio-type estimators mentioned.

# Acknowledgments

The authors are grateful to the Editor-in-Chief and two referees for providing valuable comments on an earlier draft of the paper. The paper is supported by

National Natural Science Foundation of China (11161031,11361036), Natural Science Foundation of Inner Mongolia (2013MS0108) and Doctoral Fund of Ministry of Education of China (20131514110005).

[Recibido: octubre de 2013 — Aceptado: abril de 2014]

## References

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

Asok, C. & Sukhatme, B. V. (1976), 'On sampford's procedure of unequal probability sampling without replacement', *Journal of the American Statistical Association* **71**(365), 912–918.

Bondesson, L. & Grafström, A. (2011), 'An extension of Sampford's method for unequal probability sampling', *Scandinavian Journal of Statistics* **38**(2), 377–392.

Bondesson, L. & Thorburn, L. D. (2008), 'A list sequential sampling method suitable for real-time sampling', *Scandinavian Journal of Statistics* **35**(3), 466–483.

Bondesson, L., Traat, I. & Lundqvist, A. (2006), 'Pareto sampling versus conditional Poisson and Sampford sampling', *Scandinavian Journal of Statistics* **33**(4), 699–720.

Brewer, K. R. W. (1963), 'A model of systematic sampling with unequal probability', *Australian and New Zealand Journal of Statistics* **5**(1), 5–13.

Durbin, J. (1967), 'Design of multi-stage surveys for the estimation of sampling errors', *Journal of the Royal Statistical Society. Series C: Applied Statistics* **16**(2), 152–164.

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

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

Hájek, J. (1981), *Sampling from a Finite Population*, Marcel Dekker, New York.

Kadilar, C. & Cingi, H. (2004), 'Ratio estimators in simple random sampling', *Applied Mathematics and Computation* **151**(3), 893–902.

Laitila, T. & Olofsson, J. (2011), 'A two-phase sampling scheme and πps designs', *Journal of Statistical Planning and Inference* **141**(5), 1646–1654.

Olofsson, J. (2011), 'Algorithms to find exact inclusion probabilities for 2pπps sampling designs', *Lithuanian Mathematical Journal* **51**(3), 425–439.

Rośen, B. (1997*a*), 'Asymptotic theory for order sampling', *Journal of Statistical Planning and Inference* **62**(2), 135–158.

Rośen, B. (1997*b*), 'On sampling with probability proportional to size', *Journal of Statistical Planning and Inference* **62**(2), 159–191.

Sampford, M. R. (1967), 'On sampling without replacement with unequal probabilities of selection', *Biometrika* **54**(3-4), 499–513.

Sen, A. R. (1953), 'On the estimate of variance in sampling with varying probabilities', *Journal of the Indian Society of Agricultural Statistics* **5**, 119–127.

Singh, M. P. (1967), 'Multivariate product method of estimation for finite populations', *Journal of the Indian Society of Agricultural Statistics* **31**, 375–378.

Tillé, Y. (2006), *Sampling Algorithms*, Springer, New York.

Zaizai, Y., Miaomiao, L. & Yalu, Y. (2013), 'An efficient non-rejective implementation of the πps sampling designs', *Journal of Applied Statistics* **40**(4), 870–886.

# Appendix

The derivation of the recursive formula is stated in this part.

## A1. Recursive Formula of the First-Order Inclusion Probabilities

**Proof of Lemma 1.** We use induction on $\nu$. Let $\nu = 0$, then $P_0^N = Pr(n_{s_0} = 0) = Pr\{\sum_{\alpha=1}^N I_\alpha = 0\} = Pr\{I_k = 0, \sum_{\alpha=1,\alpha\neq k}^N I_\alpha = 0\} = (1-p_k)Pr(n_{s_0}^{-k} = 0)$. Hence $Pr(n_{s_0}^{-k} = 0) = \mu_k P_0^N$, Lemma 1 is true for $\nu = 0$. Assume that equation (1) is true for $\nu = j < N$. Then

$$Pr(n_{s_0}^{-k} = j) = \mu_k \sum_{i=0}^{j} (-1)^{j-i} \big(\frac{p_k}{1-p_k}\big)^{j-i} P_i^N$$

Now, let $\nu = j+1 \leq N$. Then $P_{j+1}^N = Pr(n_{s_0} = j+1) = p_k Pr(n_{s_0}^{-k} = j) + (1-p_k)Pr(n_{s_0}^{-k} = j+1)$. By solving for $Pr(n_{s_0}^{-k} = j+1)$ and substituting in the expression above for $Pr(n_{s_0}^{-k} = j)$, we can get that

$$Pr(n_{s_0}^{-k} = j+1) = \mu_k \sum_{i=0}^{j+1} (-1)^{j+1-i} \big(\frac{p_k}{1-p_k}\big)^{j+1-i} P_i^N$$

□

**Proof of Lemma 2.** By applying Lemma 1 to the reduced population $U - \{k\}$, we can get that

$$Pr(n_{s_0}^{-kl} = \nu) = \mu_l \sum_{j=0}^{\nu} (-1)^{\nu-j} \big(\frac{p_l}{1-p_l}\big)^{\nu-j} Pr(n_{s_0}^{-k} = j)$$

Again, by substituting the expression for $Pr(n_{s_0}^{-k} = j)$ given by Lemma 1. □

**Proof of Theorem 1.** Firstly we note that

$$\pi_k = Pr(k \in s) = Pr(k \in s, n_{s_0} < n) + Pr(k \in s, n_{s_0} \geq n) \qquad (A1)$$

The first factor on the right of equation (A1) equals

$$Pr(k \in s, n_{s_0} = 0) + \sum_{\nu=1}^{n-1} [Pr(k \in s, I_k = 1, n_{s_0} = \nu) + Pr(k \in s, I_k = 0, n_{s_0} = \nu)],$$

where $Pr(k \in s, n_{s_0} = 0) = \dfrac{n}{N} Pr(n_{s_0} = 0) = \dfrac{n}{N}(1-p_k) Pr(n_{s_0}^{-k} = 0)$.

When $1 \leq \nu \leq n-1$,

$$\begin{aligned} Pr(k \in s, n_{s_0} = \nu) &= Pr(k \in s, I_k = 1, n_{s_0} = \nu) + Pr(k \in s, I_k = 0, n_{s_0} = \nu) \\ &= p_k \cdot Pr(n_{s_0}^{-k} = \nu - 1) + (1 - p_k) \cdot \frac{n-\nu}{N-\nu} \cdot Pr(n_{s_0}^{-k} = \nu), \end{aligned}$$

where $n_{s_0}^{-k} = \sum_{j \neq k}^{N} I_j$. The last equality follows from the fact that $I_k$ and $n_{s_0}^{-k}$ are independent. After some simple algebraic operation, it follows that

$$\begin{aligned} &Pr(k \in s, n_{s_0} < n) \\ &= \tfrac{n}{N}(1-p_k) Pr(n_{s_0}^{-k} = 0) + \textstyle\sum_{\nu=1}^{n-1} \big[p_k Pr(n_{s_0}^{-k} = \nu - 1) \\ &+ (1-p_k) \tfrac{n-\nu}{N-\nu} Pr\big(n_{s_0}^{-k} = \nu\big)\big] \end{aligned} \qquad (A2)$$

With the same notation and technique, we also derive that the second factor on the right of equation (A1) corresponds to

$$Pr(k \in s, n_{s_0} \geq n) = \sum_{\nu=n-1}^{N-1} p_k \cdot \frac{n}{\nu+1} \cdot Pr(n_{s_0}^{-k} = \nu) \qquad (A3)$$

By substituting (A3) and (A2) in the equation (A1) and some algebraic operations, the first-order inclusion probabilities can then be expressed as

$$\pi_k = \sum_{\nu=0}^{n-1} \Big[\frac{(N-n)p_k + (n-\nu)}{N-\nu} \cdot Pr(n_{s_0}^{-k} = \nu)\Big] + \sum_{\nu=n}^{N-1} \frac{np_k}{\nu+1} \cdot Pr(n_{s_0}^{-k} = \nu) \quad (A4)$$

By applying Lemma 1 to $Pr(n_{s_0}^{-k} = \nu)$ of equation (A4), we can get Theorem 1. □

## A2. Recursive Formula of the Second-Order Inclusion Probabilities

**Proof of Lemma 2.** The second-order inclusion probabilities can be written as

$$\pi_{ij} = Pr(i \in s, j \in s, n_{s_0} < n) + Pr(i \in s, j \in s, n_{s_0} \geq n) \qquad (A5)$$

The first expression on the right of equation (A5) equals

$$Pr(i \in s, j \in s, n_{s_0} = 0) + Pr(i \in s, j \in s, n_{s_0} = 1) + \sum_{\nu=2}^{n-1} Pr(i \in s, j \in s, n_{s_0} = \nu),$$

where $Pr(i \in s, j \in s, n_{s_0} = 0) = q_i q_j \frac{n(n-1)}{N(N-1)} Pr\big(n_{s_0}^{-ij} = 0\big)$ and

$$\begin{aligned} & Pr(i \in s, j \in s, n_{s_0} = 1) \\ = \; & Pr(i \in s, j \in s, I_i = 0, I_j = 0, n_{s_0} = 1) \\ & + Pr(i \in s, j \in s, I_i = 1, I_j = 0, n_{s_0} = 1) \\ & + Pr(i \in s, j \in s, I_i = 0, I_j = 1, n_{s_0} = 1) \\ = \; & q_i q_j \frac{(n-1)(n-2)}{(N-1)(N-2)} Pr(n_{s_0}^{-ij} = 1) + (p_i q_j + q_i p_j) \frac{(n-1)}{(N-1)} Pr\big(n_{s_0}^{-ij} = 0\big) \end{aligned}$$

When $2 \leq \nu \leq n-1$,

$$\begin{aligned} & Pr(i \in s, j \in s, n_{s_0} = \nu) \\ = \; & Pr(i \in s, j \in s, I_i = 0, I_j = 0, n_{s_0} = \nu) \\ & + Pr(i \in s, j \in s, I_i = 1, I_j = 0, n_{s_0} = \nu) \\ & + Pr(i \in s, j \in s, I_i = 0, I_j = 1, n_{s_0} = \nu) \\ & + Pr(i \in s, j \in s, I_i = 1, I_j = 1, n_{s_0} = \nu) \\ = \; & q_i q_j \frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)} P(n_{s_0}^{-ij} = \nu) + (p_i q_j + q_i p_j) \frac{n-\nu}{N-\nu} Pr(n_{s_0}^{-ij} = \nu - 1) \\ & + p_i p_j Pr(n_{s_0}^{-ij} = \nu - 2) \end{aligned}$$

The second factor on the right of equation (A5) corresponds to

$$\sum_{\nu=n}^{N} Pr(i \in s, j \in s, I_i = 1, I_j = 1, n_{s_0} = \nu) = \sum_{\nu=n}^{N} p_i \cdot p_j \cdot \frac{n(n-1)}{\nu(\nu-1)} \cdot Pr(n_{s_0}^{-ij} = \nu - 2)$$

On substituting the expressions above in equation (A5), the $\pi_{ij}$ becomes

$$\begin{aligned} \pi_{ij} = \sum_{\nu=0}^{n-2} \Big[ & (1-p_i)(1-p_j) \frac{(n-\nu)(n-\nu-1)}{(N-\nu)(N-\nu-1)} \\ & + (p_i + p_j - 2p_i p_j) \frac{n-\nu-1}{N-\nu-1} + p_i p_j \Big] Pr\big(n_{s_0}^{-ij} = \nu\big) \end{aligned}$$

$$+\sum_{\nu=n-1}^{N-2} p_i p_j \frac{n(n-1)}{(\nu+2)(\nu+1)} Pr\big(n_{s_0}^{-ij}=\nu\big) \qquad (A6)$$

By using Lemma 2 to $Pr(n_{s_0}^{-ij}=\nu)$ of equation (A6), we may derive Theorem 2. □