Metrika (1991) 38:335–343

Some Order Relations Between Selection and Inclusion Probabilities for PPSWOR Sampling Scheme

By T. J. Rao[1], S. Sengupta[2] and B. K. Sinha[1]

Abstract: In this paper we study some order relations between the selection and the inclusion probabilities for PPSWOR Sampling Scheme. We also establish some interesting bounds on the inclusion probabilities in terms of the selection probabilities.

1 Introduction

Consider a finite population $U = \{1, \ldots, N\}$ of N identifiable units and a positive-valued size measure X taking value X_i on unit i. We denote by p_i the normed size measure $X_i / \sum_{i=1}^{N} x_i$ and by $\boldsymbol{p} = (p_1, \ldots, p_N)$ the normed size vector. A sample of size n is selected from the population using a probability proportional to size without replacement (PPSWOR) sampling design based on $\boldsymbol{p}$. For an arbitrary subset s of U let $\pi(s)$ denote the probability of including s in the sample. For $s = \{i\}$, $\pi(s)$ will be denoted by π_i. In many practical survey problems it is of interest to control the values of π_i (and also of $\pi\{i,j\}$) in order to get stable estimates or to obtain certain preferred samples. It is thus important to find out how the values of π_i and π_j are related for a given relationship between p_i and p_j. It is also possible that a transformation on the available size measure or an altogether different choice of a new size measure would lead to a more suitable choice of π_i values, say π_i', under the PPSWOR scheme. In this paper we consider some questions relating to the behaviour of π_i values for a given $\boldsymbol{p}$ as well as the behaviour of π_i' and π_i vis-a-vis the relationship between p_i' and p_i, where

[1] T. J. Rao and B. K. Sinha, Stat. Math. Division, Indian Statistical Institute, 203, B. T. Road, Calcutta – 700035, India.
[2] S. Sengupta, Department of Statistics, Calcutta University 35, Ballygunge Circular Road, Calcutta – 700019, India.

$p_i' = X_i'/\sum_{i=1}^{N} X_i', X_i'$ being the value on the unit i for the alternative choice of the size measure. In the process we also establish some interesting bounds on π_i's in terms of p_i's.

2 Main Results

For PPSWOR sampling scheme based on an initial selection probability vector p, the probability of selecting an ordered sample $(i_1, \ldots, i_n)$, $1 \leq i_1 \neq \ldots \neq i_n \leq N$, is given by

$$p(i_1, \ldots, i_n) = p_{i_1} \ldots p_{i_n} (1-p_{i_1})^{-1} \ldots (1-p_{i_1} - \ldots - p_{i_{n-1}})^{-1} . \tag{2.1}$$

The probability of selecting an unordered sample (subset of size n) s_n is obtained by summing the probabilities of selection of $n!$ ordered samples given by the $n!$ permutations of the elements of s_n. Andreatta and Kaufman (1986) in their Theorem 4.1 obtained an interesting and useful integral representation of $p(s_n)$ given by

$$\begin{aligned} p(s_n) &= \left(\sum_{i=1}^{N} X_i - \sum_{i \in s_n} X_i \right) \int_0^{\infty} e^{-\lambda(\sum_{i=1}^{N} X_i - \sum_{i \in s_n} X_i)} \prod_{i \in s_n} (1 - e^{-\lambda X_i}) d\lambda \\ &= \left(1 - \sum_{i \in s_n} p_i \right) \int_0^{1} t^{-\sum_{i \in s_n} p_i} \prod_{i \in s_n} (1 - t^{p_i}) dt \end{aligned} \tag{2.2}$$

(on substituting $t = e^{-\lambda \sum_{i=1}^{N} X_i}$).

For a given p, we first prove certain inequalities connecting π_i's and corresponding p_i's. We first have an order relation involving π_i's and p_i's. More generally we prove the following.

Theorem 2.1: Let $i, j \in U$, $i \neq j$, $s_0 \subset U \backslash \{i, j\}$ with $0 \leq l = |s_0| \leq n-1$, $s_1 = s_0 + \{i\}$ and $s_2 = s_0 + \{j\}$. Then $\pi(s_1) \geq \pi(s_2)$ according to $p_i \geq p_j$.

Proof: Consider first $l = n-1$ and assume without loss of generality $i = 1$, $j = 2$. Then $s_1 = \{i_1, i_2, \ldots, i_n\}$, $s_2 = \{j_1, j_2, \ldots, j_n\}$ with $i_1 = 1$, $j_1 = 2$, $i_v = j_v$, $2 \leq v \leq n$ and

$|s_1| = |s_2| = n$. We then have $\pi(s_1) - \pi(s_2) = \sum_{\tau \in T_n} c_\tau [H_\tau(p_1) - H_\tau(p_2)]$, where T_n is the set of all permutations τ of $\{1, \ldots, n\}$,

$$c_\tau = \prod_{v=1}^{v_0} \left[1 - \sum_{\mu=1}^{v-1} p_\mu^* \right]^{-1} \prod_{\substack{j=1 \\ j \neq v_0}}^{n} p_j^* ,$$

$$H_\tau(x) = x \cdot \prod_{v=v_0+1}^{n} [d_v(\tau) - x]^{-1} , \quad d_v(\tau) = 1 - \sum_{\substack{\mu=1 \\ \mu \neq v_0}}^{v-1} p_\mu^* ,$$

$v_0 = \tau^{-1}(1)$ and $p_v^* = p_{i_{\tau(v)}}$ for $v \neq v_0$.

Since $c_\tau > 0$ and $H_\tau(x)$ is strictly increasing for $0 < x < \min_{v_0+1 \leq v \leq n} d_v(\tau)$ the assertion follows for $l = n - 1$.

For general $l (0 \leq l \leq n-1)$ the assertion follows from the first case by noting that $\pi(s_1) - \pi(s_2) = \sum_{s \in S^*} [\pi(s_1 Us) - \pi(s_2 Us)]$, where S^* consists of all subsets s from $U \backslash (s_0 U \{i,j\})$ containing exactly $n - l - 1$ elements.

As corollary we obtain

Corollary 2.1: If for any i and $j, p_i \geq p_j$ then $\pi_i \geq \pi_j$ and conversely.

Remark 2.1: Such order relation between π_i's and p_i's is trivially true for all πPS sampling schemes, where $\pi_i = np_i$. This is also obvious for Midzuno-Sen (Midzuno (1952), Sen (1952)) sampling scheme, for in this case

$$\pi_i = \frac{n-1}{N-1} + \frac{N-n}{N-1} p_i .$$

Rao (1961) and Seth (1966) considered the sampling scheme, where the first two units are drawn by PPSWOR and the remaining $(n-2)$ units of the sample by simple random sampling without replacement (SRSWOR). For this scheme, we have

$$\pi_i = \pi_i(2) \frac{N-n}{N-2} + \frac{n-2}{N-2}$$

where $\pi_i(2)$ is the inclusion probability of unit i in the first two draws. From Corollary 2.1, it follows that if $p_i \geq p_j$ then $\pi_i(2) \geq \pi_j(2)$ and hence $\pi_i \geq \pi_j$.

For the case of SRSWOR, we have $p_i = 1/N$, $\pi_i = n/N \forall i$. When we deviate from the SRSWOR probability and consider a general $\boldsymbol{p}$, we shall find out how the corresponding π_i's behave. We first establish the following theorem which provides some simple bounds on π_i's in terms of $\min(p_i)$ and $\max(p_i)$.

Theorem 2.2: Let $i_1, i_2 \in U$ with $p_{i_1} = \min_{1 \le i \le N} p_i$ and $p_{i_2} = \max_{1 \le i \le N} p_i$ and let $p_i T_r(i)$ denote the probability of getting unit i in the rth draw, $1 \le r \le n$, $1 \le i \le N$. Then

(a) $T_r(i_1) \le T_{r+1}(i_1)$

(b) $T_r(i_2) \ge T_{r+1}(i_2)$, $1 \le r \le n$

(c) $np_{i_1} \le \pi_{i_1} \le n/N$ and $n/N \le \pi_{i_2} \le np_{i_2}$

(d) $np_{i_1} \le \pi_i \le np_{i_2}$, $i \in U \backslash \{i_1, i_2\}$.

Proof: We can write

$$T_{r+1}(i_1) = \sum_{s_r \in S_r} p(s_r) \left(1 - \sum_{i \in s_r} p_i\right)^{-1} ,$$

where $p(s_r)$ denotes the probability of obtaining a PPSWOR (N, r) subset s_r and S_r denotes the set of all subsets of $U \backslash \{i_1\}$ containing exactly r units. Now

$$\begin{aligned}
T_{r+1}(i_1) &= \sum_{s_r \in S_r} \left(1 - \sum_{i \in s_r} p_i\right)^{-1} \sum_{s_{r-1} \subset s_r} p(s_{r-1}) \left(1 - \sum_{i \in s_{r-1}} p_i\right)^{-1} \sum_{i \in s_r \backslash s_{r-1}} p_i \\
&= \sum_{s_{r-1} \in S_{r-1}} p(s_{r-1}) \left(1 - \sum_{i \in s_{r-1}} p_i\right)^{-1} \sum_{\substack{s_r \supset s_{r-1} \\ s_r \in S_r}} \left(1 - \sum_{i \in s_r} p_i\right)^{-1} \sum_{i \in s_r \backslash s_{r-1}} p_i \\
&\ge \sum_{s_{r-1} \in S_{r-1}} p(s_{r-1}) \left(1 - \sum_{i \in s_{r-1}} p_i\right)^{-1} \left(1 - p_{i_1} - \sum_{i \in s_{r-1}} p_i\right)^{-1} \\
&\quad \times \sum_{\substack{s_r \supset s_{r-1} \\ s_r \in S_r}} \sum_{i \in S_r \backslash S_{r-1}} p_i \\
&= \sum_{s_{r-1} \in S_{r-1}} p(s_{r-1}) \left(1 - \sum_{i \in s_{r-1}} p_i\right)^{-1} = T_r(i_1) ,
\end{aligned}$$

which proves (a). The part (b) follows by similar arguments.

The part of the inequalities in (c) with n/N as a bound is easily established by contradiction. Let, if possible, $\pi_{i_1} > n/N$. Then, by Corollary 2.1, $\pi_i > n/N \quad \forall i$, which contradicts the relation that $\sum_1^N \pi_i = n$. Hence, we must have $\pi_{i_1} \leq n/N$. Similarly, it can be shown that $\pi_{i_2} \geq n/N$.

To prove the other part of the inequalities in (c) we note that $\pi_i = p_i \sum_{r=1}^{n} T_r(i)$, $T_1(i) = 1$, whence the assertions follow by (a) and (b). The part (d) follows from (c) and Corollary 2.1.

Consider now a simple deviation from the SRSWOR given by the initial selection probability vector

$$p = (p^{(1)}, \ldots, p^{(1)}, 1/N, \ldots, 1/N, p^{(2)}, \ldots, p^{(2)}) \ ,$$

where $p^{(1)} < 1/N < p^{(2)}$. This situation may occur in practice when one has three different types of units homogeneous within each type and one wishes to select them with three different types of probabilities say $p^{(1)}$, $1/N$ and $p^{(2)}$. If $\pi^{(i)}$ denotes the inclusion probability of ith type, $i = 1,2,3$, it follows from (c) of Theorem 2.2 that $\pi^{(1)} \geq np^{(1)}$ and $\pi^{(3)} \leq np^{(2)}$. However, it does not follow that $\pi^{(2)} \geq$ or $\leq n/N$. It may thus be of interest to compare the values of $\pi^{(2)}$ with n/N and more generally, the value of π_i with np_i for a general p.

Towards this, we establish the following theorem giving lower bounds on π_i's in terms of corresponding p_i's.

Theorem 2.3: For a given $p = (p_1, \ldots, p_N)$,

$$\pi_i \geq p_i c(p_i) \ , \quad 1 \leq i \leq N \ ,$$

where $c(p_i)$ is the value of π_i/p_i based on

$$p^{(i)} = \left(\frac{1-p_i}{N-1}, \ldots, \frac{1-p_i}{N-1}, p_i, \frac{1-p_i}{N-1}, \ldots, \frac{1-p_i}{N-1} \right).$$

Proof: Without loss of generality, take $i = 1$. We then have,

$$\pi_1 = p_1 \left(1 + \sum_{r=2}^{n} T_r(1) \right) = p_1 (1 + g(p_2, \ldots, p_N)) \ , \quad \text{say} \tag{2.3}$$

where $T_r(1)$ is defined as in Theorem 2.2. Now, by Lemma A.2 of the Appendix, $g(p_2, \ldots, p_N)$ and hence, π_1/p_1 is minimum subject to $\sum_2^N p_i = 1 - p_1$ when

$p_2 = \ldots = p_N = (1-p_1)/N-1$. This completes the proof of the theorem.

It is easy to verify that $c(p_i)$ is decreasing in p_i and that for $p_i = 1/N$, $c(p_i) = n$. From these, we immediately have the following.

Corollary 2.2: If $p_i \leq p_{i0}$, then $\pi_i \geq p_i c(p_{i0})$.

Corollary 2.3: If $p_i = 1/N$, then $\pi_i \geq n/N$.

Corollary 2.4: If $p_i \leq 1/N$, then $\pi_i \geq np_i$.

Remark 2.2: For $p_i > 1/N$, it is not, however, necessarily true that $\pi_i < np_i$. The following is a counter-example.

Example 2.1: $N = 3$. $p = (0.01, 0.34, 0.65)$. Let $n = 2$. Here $p_2 > 1/N = 0.3333$. But $\pi_2 = 0.9748 > 2p_2 = 0.68$.

Motivated by these results, we shall now ask the question whether for two initial selection probability vectors $\boldsymbol{p}$ and $\boldsymbol{p}'$, the corresponding vector of inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ and $\boldsymbol{\pi}' = (\pi'_1, \ldots, \pi'_N)$ are related in the same way as $\boldsymbol{p}$ and $\boldsymbol{p}'$ are.

The following example shows that $p_i < p'_i$ does not necessarily imply that $\pi_i < \pi'_i$ even in the case where p'_i is obtained as a simple transformation from p_i, namely $p'_i = \alpha + \beta p_i$.

Example 2.2: Let $N = 3$, $n = 2$, and $p'_i = 0.1/3 + 0.9p_i$. For $\boldsymbol{p} = (0.1, 0.2, 0.7)$, $\boldsymbol{p}' = (0.37/3, 0.64/3, 1.99/3)$. Here $p_2 < p'_2$ but $\pi_2 = 0.6889 > \pi'_2 = 0.6637$.

Remark 2.3: In this connection we may note that such relation between $\boldsymbol{p}$ and $\boldsymbol{\pi}$ is true if the scheme used is a πPS one or the scheme due to Midzuno-Sen (1952) (see Remark 2.1).

Finally, we investigate whether the vector of inclusion probabilities $\boldsymbol{\pi}$ is isotonic in $\boldsymbol{p}$. The specific question can be posed as follows. Let $\boldsymbol{p}$ and $\boldsymbol{p}'$ be two normed size vectors satisfying $\boldsymbol{p} < \boldsymbol{p}'$ and let $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$ be the corresponding vectors of inclusion probabilities underlying the PPSWOR desings each of size n. Is it then true that $\boldsymbol{\pi} < \boldsymbol{\pi}'$? Curiously enough, in general, this does not hold. A counter-example is obtained by taking

Example 2.3: $N=3$, $n=2$, $p=(0.1,0.3,0.6)$ and $p'=(0.32/3,1.3/3,1.38/3)$, resulting in $\pi=(0.2929,0.7833,0.9238)$ and $\pi'=(0.2791,0.8542,0.8667)$. It is readily seen that $p<p'$ while $\pi>\pi'$.

However, the following theorem shows that if we compare any initial probability vector p and the corresponding π with the SRSWOR probabilities, namely $p_0=(1/N,\ldots,1/N)$ and $\pi_0=(n/N,\ldots,n/N)$, we always have $p<p_0$ and $\pi<\pi_0$.

Theorem 2.4: For every normed size vector $p=(p_1,\ldots,p_N)$, $p_1\leq p_2\leq\ldots\leq p_N$ we have

$$\text{(i)}\ \sum_{j=1}^{t}\pi_j\leq tn/N\quad\forall t\geq 1\quad\text{and}\quad\text{(ii)}\ \sum_{j=1}^{t}p_j\leq t/N\quad\forall t\geq 1\ .$$

Proof: For $t=1$, (i) follows by (c) of Theorem 2.2. Suppose now (i) holds for $t=t$, but does not hold for $t=t+1$. Then we must have $\pi_{t+1}>n/N$ implying by Corollary 2.1, $\pi_j>n/N\quad\forall j=t+2,\ldots,N$. This gives $\sum_{j=1}^{N}\pi_j=\sum_{j=1}^{t+1}\pi_j+\sum_{j=t+2}^{N}\pi_j>\frac{(t+1)n}{N}+\frac{(N-t-1)n}{N}=n$, which contradicts the relation that $\sum_{j=1}^{N}\pi_j=n$. Hence, if (i) holds for t then it also holds for $t+1$. As (i) holds for $t=1$, it holds $\forall t\geq 1$.

It can be established by similar arguments (replacing π_j by p_j and tn/N etc. by t/N etc.) that (ii) holds.

Appendix

Lemma A.1: Let $a>1$, $c>0$ and integers r,k with $1\leq r\leq k$ be given. Define $X=\left\{x=(x_1,\ldots,x_k)\in R^k: x_i\geq 0,\ \sum_{i=1}^{k}x_i=c\right\}$ and $f_p(a,x)=\sum_{l=1}^{r}\sum_{s\in S_l}\prod_{i\in s}(a^{x_i}-1)$ for $x\in X$, where S_l denotes the set of all subsets of $\{1,\ldots,k\}$ containing exactly l elements. Then for $x^*\in X$ it holds that $f(x^*)=\min_{x\in X}f(x)$ iff $x_i^*=c/k$, $1\leq i\leq k$.

Proof: Since X is a compact subset of R^k and f is continuous on X there exists at least one $x^*\in X$ with $f(x^*)=\min_{x\in X}f(x)$. Suppose now that one component of

x^* is not equal to c/k. Then there are $i_1, i_2 \in \{1, \ldots, k\}$ such that $x^*_{i_1} < c/k < x^*_{i_2}$. Let now $\varepsilon > 0$ with $x^*_{i_1} + \varepsilon \leq c/k \leq x^*_{i_2} - \varepsilon$ and define

$$\hat{x}_i = \begin{cases} x^*_i , & \text{if } i \neq i_1, i_2 \\ x^*_{i_1} + \varepsilon , & \text{if } i = i_1 \\ x^*_{i_2} - \varepsilon , & \text{if } i = i_2 \end{cases}$$

Of course $x \in X$. Now it is easy to show

$$f(x^*) - f(\hat{x}) = b \cdot a^{-\varepsilon}(a^{\varepsilon} - 1) \cdot (a^{x^*_{i_2}} - a^{x^*_{i_1} + \varepsilon}) ,$$

where $b = \sum_{s \in S^*} \prod_{i \in s} (a^{x^*_i} - 1)$ and S^* is the set of all subsets of $\{1, \ldots, k\} \setminus \{i_1, i_2\}$ containing exactly $r-1$ elements. Because of $b > 0$, $\varepsilon > 0$, $a > 1$ and $x^*_{i_2} > x^*_{i_1} + \varepsilon$ one would get $f(x^*) > f(\hat{x})$ in contradiction to $f(x^*) = \min_{x \in X} f(x)$. This completes the proof of the lemma.

Lemma A.2: For PPSWOR (N, n) sampling scheme with an initial selection probability vector $p = (p_1, \ldots, p_N)$, let $g(p_2, \ldots, p_N)$ be defined as in (2.3). Then g is minimum subject to $\sum_{i=2}^{N} p_i = 1 - p_1$ when $p_i = (1 - p_1)/(N-1) \quad \forall i = 2, \ldots, N$.

Proof: Observe that

$$g(p_2, \ldots, p_N) = \sum_{r=1}^{n-1} \sum_{s_r \in S_r} p(s_r) \left(1 - \sum_{i \in s_r} p_i\right)^{-1} ,$$

where $p(s_r)$ denotes the probability of a PPSWOR (N, r) subset s_r and S_r denotes the set of all subsets of $U \setminus \{1\}$ containing exactly r units. Using the integral representation (2.2), we have

$$g(p_2, \ldots, p_N) = \sum_{r=1}^{n-1} \sum_{s_r \in S_r} \int_0^1 \prod_{i \in s_r} (t^{-p_i} - 1)\, dt = \int_0^1 f_{n-1}(1/t, p_2, \ldots, p_N)\, dt$$

in the notation of the function f defined in Lemma A.1. By Lemma A.1, it now follows that $f_{n-1}(1/t, p_2, \ldots, p_n)$ is minimum when $p_i = \frac{1-p_1}{N-1} \quad \forall 2 \leq i \leq N$, for every $t \in (0, 1)$. Hence, the lemma follows.

Acknowledgement: The authors like to record their sincere thanks to the referee for suggesting the present proof of Lemma A.1 and for many other suggestions which considerably improved the presentation of the results.

The authors also thank Dr. S. Das Gupta and Dr. J.C. Gupta for their comments which led to Example 2.3 and Theorem 2.4.

References

1. Andreatta G, Kaufman GM (1986) Estimation of finite population properties when sampling is without replacement and proportional to magnitude. J Amer Statist Assoc 81:657–666
2. Midzuno H (1952) On the sampling system with probability proportional to the sum of sizes. Ann Inst Stat Math 3:99–108
3. Rao JNK (1961) On the estimate of variance in unequal probability sampling. Ann Inst Stat Math 15:67–72
4. Sen AR (1952) Present status of probability sampling and use in estimation of farm characteristics (abstract). Econometrica 20:103
5. Seth GR (1966) On estimators of variance of estimate of population total in varying probabilities. J Ind Soc Agr Stat 18:52–56

Received 29 November 1989
Revised version 7 September 1990