A Two-phase Sampling Scheme and $\pi ps$ Designs

Thomas Laitila$^1$, Jens Olofsson$^2$

Abstract

In this paper a two-phase approach that gives a fixed sample size and unequal inclusion probabilities is presented. It is proposed that population parameters can be estimated by using the theory for probability proportional to size sampling. It is shown, by means of simulation, that the associated estimators work well with respect to empirical bias and precision. The scheme corresponds to a two-phase design allowing for standard inference.

Key words: $\pi ps$ sampling design, simulation, estimation.

1 Introduction

Usually a sample survey is taken in order to estimate parameters of a finite population $U = \{1, 2, \ldots, k, \ldots, N\}$ such as totals or functions thereof. In cases when there is only one population parameter to be estimated, say

$$t = \sum_U y_k$$

that is a population total of variable $y$ and where $y_k$ is the value of the study variable $y$ for the $k$th element in $U$, the parameter could be estimated by using

$$\hat{t} = \sum_s w_k y_k$$

where $w_k$ is the design weight and $s$ the sample. If $1/\pi_k$ is used as design weights, then (2) is the well-known $\pi$-estimator (Särndal, Swensson and Wretman, 1992) and the estimator is unbiased (Horvitz and Thompson, 1952).

It is often possible to increase the precision of the survey by using a sampling design with unequal inclusion probabilities rather than a design where the inclusion probabilities are equal for all elements in the population, that is a design with inclusion probabilities $\pi_k \neq \pi_l$.

A simple without replacement design with unequal inclusion probabilities is the Poisson (PO) design. The optimal choice of the first-order inclusion probabilities would be $\pi_k \propto y_k$ for all $k \in U$, since the variation of (2) would solely then be due to the variation in the sample size, $n_s$ (see for example Särndal et alia, 1992). If the design has a fixed size, the variation would be zero.

$^1$Thomas Laitila, Statistics Sweden and Örebro University, Department of Statistics, Örebro University, SE-701 82 Örebro, Sweden, thomas.laitila@oru.se

$^2$Jens Olofsson, Department of Statistics, Örebro University, SE-701 82 Örebro, Sweden, jens.olofsson@oru.se
Although $\pi_k \propto y_k$ is not possible, since it requires complete knowledge of the study variable $y_k$ for each $k \in U$ on beforehand and hence, no survey would be required. In some cases, on the other hand, there exists auxiliary information in shape of some size variable $x$ which covaries positively with the study variable $y$. By taking the auxiliary information into account and by letting

$$\pi_k \propto x_k$$

it is possible to reduce the variance of (2) in comparison to for example using a simple random sampling (SI) design. Without replacement designs complying with the $\pi$ps rule (3) are said to be $\pi$ps designs. If the design in question also has a fixed size the target inclusion probabilities are given by

$$\lambda_k = \frac{nx_k}{N\bar{x}_U}$$

where $n$ is the predetermined sample size and $N$ the population size.

A major drawback with PO designs complying with the $\pi$ps rule is the randomness of the sample size $n_s$. In the literature different fixed size sampling designs that comply with (3) at least approximately or for a subset of the sample $s$ has been proposed, see Brewer and Hanif (1983) for an overview. Sampford (1967) suggests a rejective design which yields exact first-order inclusion probabilities complying with (3), whereas the conditional PO design suggested by Hajek (1964) only yield inclusion probabilities that approximately comply with (3). Designs which incorporate order sampling have also been suggested (Saavedra, 1995, Kröger, Särndal and Teikari 2003). Out of these latter designs perhaps the most well-known is the Pareto $\pi$ps design, see Rosén (1997a,b).

In this paper a sampling scheme is proposed for fixed size designs with unequal inclusion probabilities. The scheme is neither rejective nor does it utilize order sampling. It is based on a two-phase scheme and uses any without replacement design with unequal inclusion probabilities in the first phase and an SI design in the second. In particular the scheme is suggested for generating fixed size $\pi$ps samples. The factual first-order inclusion probabilities of the proposed scheme are approximately equal to target inclusion probabilities (4). This suggests to treat the sample as a true $\pi$ps sample using (2) for estimation with the reciprocal of the target inclusion probabilities as design weights. The scheme is easy to implement and has an interesting feature: it corresponds to a two-phase design allowing for standard inference.

2 A Two-phase Fixed Size Sampling Design

Two-phase sampling design (or double sampling design) was first introduced by Neyman (1938), with SI design in the first phase and a stratified simple random sampling (STSI) design in the second phase. General formulas for variances and variance estimation irrespective of design in each phase were derived by Särndal and Swensson (1987). A two-phase sampling design could e.g. be used when there exists no informative sampling frame to stratify from as in Neyman (1938) or as a way of handling non-response, see
Särndal and Swensson (1987) or Särndal et alia (1992, ch. 15). In this paper the two-phase sampling design is used to generate a sample from a design with fixed size and unequal inclusion probabilities complying with (3).

2.1 The 2Pπps Design

Let \( n \) be the predetermined sample size and assume target inclusion probabilities, \( \lambda_k \), to be proportional to a size variable, \( x \), known for all \( k \in U \), that is \( \lambda_k \propto x_k \). The sampling scheme proposed is as follows:

1. Draw a sample using a PO design with \( \lambda_{ak} \propto x_k \) as parameters and with expected sample size \( \sum_U \lambda_{ak} = m \geq n \).

2. If the size of the sampled set, \( s_a \), is smaller than the predetermined sample size, that is \( n_{sa} < n \), then repeat step 1. If not, proceed to step 3.

3. From the sampled set, \( s_a \), draw a sample of size \( n \) using a SI design.

The sampling scheme proposed corresponds to a sampling design, here called the 2Pπps design, with first- and second order inclusion probabilities given by

\[
\pi_k = \pi_{ak} \mathbb{E}_{pa}\left( \frac{n}{n_{sa}} \mid k \in s_a, n_{sa} \geq n \right)
\]  
(5)

and

\[
\pi_{kl} = \pi_{akl} \mathbb{E}_{pa}\left( \frac{n(n-1)}{n_{sa}(n_{sa}-1)} \mid k \& l \in s_a, n_{sa} \geq n \right)
\]  
(6)

respectively. Here \( \pi_{ak} \) and \( \pi_{akl} \) are the factual first- and second order inclusion probabilities, respectively, in the first phase using a PO design with \( \lambda_{ak} \propto x_k \) as parameters where only samples larger than or equal to \( n \) are accepted. The factual first-order inclusion probabilities in the first phase of the 2Pπps design are given by

\[
\pi_{ak} = \lambda_{ak} \frac{\text{Pr}(n_{sa}^k \geq n - 1)}{\text{Pr}(n_{sa} \geq n)}
\]  
(7)

where \( n_{sa}^k = \sum_{U \setminus \{k\}} I_{ak} \) and

\[
I_{ak} = \begin{cases} 
1 & \text{if } k \in s_a \\
0 & \text{otherwise}. 
\end{cases}
\]  
(8)

Equation (5) shows that the approximation to the target inclusion probabilities improves with population size. Note that a different design could be used in the first step of scheme in order to obtain another set of target inclusion probabilities. For instance, Bernoulli designs gives inclusion probabilities corresponding to those of a SI design. Using a PoMix design in the first step of the scheme gives inclusion probabilities approximating those from another PoMix design.
2.2 Estimation

Based on the 2Pπps design population parameters can be estimated using standard two-phase theory, see Särndal and Swensson (1987) or Särndal et alia (1992, ch. 9). An unbiased estimator of (1) is given by

\[ \hat{t}_{2Pπ^{\star}} = \sum_s \frac{y_k}{\bar{π}} = \sum_s \frac{y_k}{\bar{π}_{ak} \bar{π}|s_a} \]  

(9)

Another option is to regard the sample as a true πps sample and use the reciprocal of the target first-order inclusion probabilities as design weights in (2), where the target probabilities are given by (4). Hence, using the 2Pπps design, an estimator of (1) is given by

\[ \hat{t}_{2Pπps,λ} = \sum_s \frac{y_k}{λ_k} \]  

(10)

Using (10) as estimator of (1) will result in some bias since, for the proposed design, \( λ_k \approx \bar{π}_k \) for all \( k \in U \).

A variance estimator for (10) has not yet been derived. However, a possible choice could be the variance estimator used for (10) when using a Pareto πps design. The variance is based on an approximation by Hajek (Hajek, 1964) and is given by (Rosén, 1997b)

\[ \hat{V}(\hat{t}) = \frac{n}{n-1} \sum_s \left( \left( \frac{y_k}{λ_k} - \frac{\sum_s y_k (1 - λ_k)}{\sum_s (1 - λ_k)} \right)^2 (λ_k - 1) \right) \]  

(11)

3 Simulation

The 2Pπps design proposed in Section 2 is here studied by means of simulation. The first-order inclusion probabilities in the PO design of the first phase are

\[ λ_{ak} = m x_k/N \bar{x}_U. \]  

(12)

Here \( m \) is set to \( \lceil \sum_U (x) / \max_k (x_k) \rceil \) since it is the largest possible expected sample size in order to avoid first-order inclusion probabilities larger or equal to one. The approximation of factual first-order inclusion probabilities in the first phase by \( λ_{ak} \) works rather well for this choice of \( m \) since the ratio in (7) tends to 1 as the expected sample size in the first phase increases. Approximating \( π_{ak} \) by \( λ_{ak} \) eases the computational efforts when using the standard two-phase theory for estimation, although \( π_{ak} \) is possible to compute exactly due to a formula implicit by Chen, Dempster and Liu (1994), see for example also Bondesson, Traat and Lundquist (2006).
3.1 Setup

The population in the simulation is a survey population which constitutes the National Travel Survey in Sweden 2005–2006. As study variable, $y$, the length, in kilometers, of a travel by car during an ordinary day has been chosen and travel time (in minutes) is used as auxiliary information, $x$. The correlation between the two variables is 0.892. Some characteristics of the population can be found in Table 1.

Table 1: Descriptives on the population

<table>
<thead>
<tr>
<th></th>
<th>$y$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>709655.1</td>
<td>897054</td>
</tr>
<tr>
<td>mean</td>
<td>52.8</td>
<td>66.8</td>
</tr>
<tr>
<td>s.e.</td>
<td>81.6</td>
<td>71.3</td>
</tr>
<tr>
<td>$Q_1$</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>26</td>
<td>45</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>60</td>
<td>85</td>
</tr>
<tr>
<td>max</td>
<td>1300</td>
<td>1200</td>
</tr>
<tr>
<td>min</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>$N$</td>
<td>13433</td>
<td>13433</td>
</tr>
</tbody>
</table>

Five different sample sizes were used, namely \{2, 5, 10, 50, 100, 500\}. The number of replicates ($R$) were 30,000 and the simulation was done with R version 2.6.1 using `set.seed(7402)` for random number generation. The estimator (10) was evaluated for the proposed design as well as for the Pareto $\pi ps$ design. For comparison (9) was also included in the simulation. Equation (11) was used as a variance estimator for the two $\pi ps$ designs.

For each of the designs three measures were computed: the empirical relative bias of the point estimator for the population total, $\text{REB}(\hat{t})$, the empirical design effect, $\text{Edeff}(p(s), t)$, and the empirical relative bias of the variance estimator for the point estimator of the population total, $\text{REB}(\hat{V}(\hat{t}))$. The empirical relative bias for parameter $\theta$ is given by

$$
\text{REB}(\hat{\theta}) = \frac{\hat{E}(\theta) - \theta}{\theta} \times 100
$$

where

$$
\hat{E}(\theta) = \frac{\sum_{r=1}^{R} \hat{\theta}_i}{R}
$$

and $\hat{\theta}_i$ is the Monte Carlo estimate of $\theta$ at the $i$th iteration. The empirical design effect for the design given by $p(s)$ with respect to the population total is given by

$$
\text{Edeff}(p(s), t) = \frac{\hat{V}_{\pi ps}(\hat{t})}{\hat{V}_{SI}(\hat{t})}
$$

(14)
where
\[ \hat{V}_{p(s)}(\hat{t}) = \frac{1}{R - 1} \sum_{r=1}^{R} (\hat{t}_i - \overline{\hat{t}})^2. \]

3.2 Results

As expected, although not formally shown in this paper, the simulation results show that, for the proposed design, \( \pi_k \approx \lambda_k \). Furthermore, the simulation shows that the relative empirical bias, see Table 2, is small for all three estimators; in absolute value the largest is about 0.12 percent (Pareto \( \pi_{ps} \) design, \( n = 5 \)). Hence, the estimator used for the proposed design seem to work well with respect to empirical relative bias. The small bias in \( \hat{t}_{2P(PO,SI),\pi^*} \) is expected since \( \lambda_{ak} \) has been used as an approximation of \( \pi_{ak} \) in order to ease the computations.

\[
\begin{array}{c|ccc}
   n & \hat{t}_{\text{PAR,} \lambda} & \hat{t}_{2P_{\pi_{ps}}, \lambda} & \hat{t}_{2P(PO,SI),\pi^*} \\
\hline
   2 & -0.102 & -0.052 & -0.083 \\
   5 & -0.124 & -0.015 & -0.030 \\
  50 & 0.045 & 0.034 & 0.027 \\
 100 & 0.023 & -0.031 & -0.051 \\
 500 & -0.001 & 0.001 & 0.006 \\
\end{array}
\]

The simulation results also show that the two \( \pi_{ps} \) designs are equally efficient in terms of precision irrespective of \( n \) for a given size of the population and expected first phase sample size, when the reciprocal of (4) are used as design weights, see Table 3. The rate at which the precision increases is the same for the two \( \pi_{ps} \) designs and the SI design as the sample size increases, but slower for the two-phase design. Hence, the gain in precision of using a two-phase design compared to a SI design decreases with an increasing sample size.

\[
\begin{array}{c|ccc}
   n & \hat{t}_{\text{PAR,} \lambda} & \hat{t}_{2P_{\pi_{ps}}, \lambda} & \hat{t}_{2P(PO,SI),\pi^*} \\
\hline
   2 & 0.107 & 0.107 & 0.108 \\
   5 & 0.107 & 0.108 & 0.111 \\
  50 & 0.107 & 0.108 & 0.132 \\
 100 & 0.106 & 0.106 & 0.154 \\
 500 & 0.102 & 0.101 & 0.346 \\
\end{array}
\]

An estimator of the variance of the \( \pi \)-estimator for the 2P\( \pi_{ps} \) design has not yet been derived, but using the variance estimator proposed seem to work well with respect to
empirical bias as shown in Table 4. Although some bias can be observed for all three designs, and the bias in the variance estimator is higher than the bias for the point estimator of the population total, it is negligible in practice.

Table 4: Relative empirical bias for $\hat{V}(\hat{t}_{\text{PAR}, \lambda})$, $\hat{V}(\hat{t}_{2P, \pi_{ps}, \lambda})$ and $\hat{V}(\hat{t}_{2P(PO,SI), \pi^*})$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\hat{V}(\hat{t}_{\text{PAR}, \lambda})$</th>
<th>$\hat{V}(\hat{t}<em>{2P, \pi</em>{ps}, \lambda})$</th>
<th>$\hat{V}(\hat{t}_{2P(PO,SI), \pi^*})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.269</td>
<td>1.112</td>
<td>1.033</td>
</tr>
<tr>
<td>5</td>
<td>0.851</td>
<td>-0.731</td>
<td>-0.344</td>
</tr>
<tr>
<td>50</td>
<td>0.953</td>
<td>-0.294</td>
<td>-0.122</td>
</tr>
<tr>
<td>100</td>
<td>0.811</td>
<td>0.764</td>
<td>0.071</td>
</tr>
<tr>
<td>500</td>
<td>0.593</td>
<td>1.209</td>
<td>0.612</td>
</tr>
</tbody>
</table>

4 Concluding Remarks

In this paper a two-phase fixed-size sampling scheme with unequal inclusion probabilities has been proposed in order to generate a sample where the first-order inclusion probabilities comply with the $\pi_{ps}$-rule (3). The proposed algorithm facilitates unbiased estimation by using the theory of two-phase sampling.

The scheme is proposed to be used as a method to generate an approximate $\pi_{ps}$ sample. Simulation results show that using the proposed design with the reciprocal of the target inclusion probabilities as design weights works well in terms of empirical bias and precision. The empirical bias is on a par with using the standard two-phase estimation, and the precision is on par with using the Pareto $\pi_{ps}$ design coupled with the $\pi$-estimator.

The proposed scheme uses basic designs and it is theoretically easier to grasp than many of the other fixed-size schemes suggested for a $\pi_{ps}$ design. It is also easy to use since the designs are implemented in most statistical softwares. For the future the properties of the inclusion probabilities for the proposed design need to be further studied, particularly in order to make a design-based variance estimation feasible, as well as the implications of different choices of $m$. 
Acknowledgments

The authors are grateful to Anton Grafström for valuable comments and suggestions which has improved this paper and the Swedish Institute for Transport and Communication Analysis for providing data.

References


