# A two-phase sampling scheme and $\pi$ps designs

Thomas Laitila[1] and Jens Olofsson[2]

[1] Statistics Sweden, Örebro University, Sweden
e-mail: thomas.laitila@esi.oru.se

[2] Örebro University, Sweden
e-mail: jens.olofsson@oru.se

## Abstract

In this paper a two-phase approach that gives a fixed sample size and unequal inclusion probabilities is presented. Population parameters can be unbiasedly estimated by using theory for two-phase sampling. An alternative is to use 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.

## 1 Introduction

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

$$t = \sum\nolimits_U y_k \tag{1}$$

i.e. a population total of variable y, where $y_k$ is the value of the study variable for the $k$th element in $U$, the parameter could be estimated by using the well-known $\pi$-estimator (Särndal, Swensson & Wretman 1992)

$$\hat{t} = \sum\nolimits_s w_k y_k \tag{2}$$

where $w_k$ is the design weight and $s$ the sample. If $1/\pi_k$ is used as design weights, then (2) is unbiased.

It is 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.

One simple family of without replacement (wor) designs with unequal inclusion probabilities is the family of Poisson (PO) designs. The optimal choice of the first-order inclusion probabilities would be by letting $\pi_k \propto y_k$ for all $k \in U$, since, if the design has fixed-size, then the variation of (2) would be zero and otherwise solely be due to the variation in the sample size, $n_s$ (see e.g. Särndal *et al* 1992). This is not possible however, since it requires complete knowledge of the study variable $y_k$ on before hand and hence, no survey would be required. In some cases, the study variable, $y$, co-varies positively with some auxiliary information in shape of some size variable $x$. By taking the auxiliary information into account and by letting

$$\pi_k \propto x_k \tag{3}$$

it is possible to reduce the variance of (2) in comparison to e.g. using a simple random sampling (SI) design. Wor designs complying with the $\pi$ps rule (3) are said to be $\pi$ps designs.

A major drawback with PO designs complying with the $\pi$ps rule is the randomness of the sample size $n_s$. In 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 & 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 designs 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 & Teikari 2003). Out of these designs perhaps the most well-known is the Pareto $\pi$ps, 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 design with any wor design in the first phase with unequal inclusion probabilities and a 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 the target probabilities. 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, viz. the scheme 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 *et al* (1992). A two-phase sampling design could e.g. be used when there exits no informative sampling frame to stratify from as in Neyman (1938) or as a way of handling non-response, see Särndal *et al* (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 2P$\pi$ps design

Let $n$ be the pre-determined sample size and assume target inclusion probabilities, $\lambda_k$, to be proportional to a size variable known for all $k \in U$, $x$, i.e $\lambda_k \propto x_k$.
The sampling scheme proposed is as follows:

1. Draw a sample using a PO design, where $\pi_{ak} \propto x_k$ with expected sample size $m \geq n$.

2. If the size of the sampled set, $s_a$, is smaller than the pre-determined sample size, i.e. $n_{s_a} < n$, then repeat step 1. If not, proceed to the next step.

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 2P$\pi$ps design, with first- and second order inclusion probabilities given by

$$\pi_k = \pi_{ak}\mathbb{E}_{p_a}\left(\frac{n}{n_{s_a}}|k \in s_a, n_{s_a} \geq n\right) \tag{4}$$

and

$$\pi_{kl} = \pi_{akl}\mathbb{E}_{p_a}\left(\frac{n(n-1)}{n_{s_a}(n_{s_a}-1)}|k\&l \in s_a, n_{s_a} \geq n\right) \tag{5}$$

respectively. Here $\pi_{ak}$ and $\pi_{akl}$ are the first- and second order inclusion probabilities, respectively, in the first phase using a PO design where $\pi_{ak} \propto x_k$. Equation (4) 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 2P$\pi$ps design population parameters can be estimated using standard two-phase theory, see Särndal *et al* (1992, ch. 9). An unbiased estimator of (1) is given by

$$\hat{t}_{2P_{PO,SI}\pi^\star} = \sum\nolimits_s \frac{y_k}{\pi^\star} = \sum\nolimits_s \frac{y_k}{\pi_{ak}\pi_{k|s_a}} \tag{6}$$

Another option is to regard the sample as a true $\pi$ps sample and use the reciprocal of the target first-order inclusion probabilities as design weights in (2), where

$$\lambda_k = \frac{n}{N\bar{x}_U} x_k \tag{7}$$

Hence, an estimator of (1) is given by

$$\hat{t}_{2P_{PO,SI}\lambda} = \sum\nolimits_s \frac{y_k}{\lambda_k} \tag{8}$$

**Remark 1** Using (8) as estimator of (1) will result in some bias since, for the proposed design, $\pi_k \approx \lambda_k$ for all $k \in U$.

**Remark 2** Rosén (1997b) uses (8) as an estimator of (1) for the Pareto $\pi$ps design and the variance estimator

$$\hat{\mathbb{V}}(\hat{t}) = \frac{n}{n-1}\sum\nolimits_s \left(\left(\frac{y_k}{\lambda_k} - \frac{\sum_s y_k(1-\lambda_k)}{\sum_s(1-\lambda_k)}\right)(\lambda_k - 1)\right) \tag{9}$$

# 3 Simulation

The 2P$\pi$ps 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

$$\pi_{ak} = mx_k/N\bar{x}_U \tag{10}$$

Here $m$ is set to $\lfloor \sum_U(x)/\max_k(x_k) \rfloor$ in order to avoid first-order inclusion probabilities larger or equal to one.

## 3.1 Setup

The the well-known MU284 population from Särndal *et al* (1992) was used as a base population with P85 as study variable $y$ and P75 as auxiliary variable $x$. The MU284 population was multiplied with 10 to the power of 3 to also have a larger population. The number of replicates in the simulation were 30 000 with R version 2.6.1. with set.seed(7402).

Table 1: Descriptives on the populations and simulation setup

| $U_i$ | $N$ | $m$ | $n$ |
|---|---|---|---|
| $U_1$ | 284 | 12 | 5 |
| $U_4$ | 284000 | 12193 | 5, 5000 |

The estimator (8) was evaluated for the proposed design as well as for the Pareto $\pi$ps design. The estimator (6) was also studied for the 2P$\pi$ps design.

## 3.2 Results

First of all, as expected, although not formally shown in this paper, the simulation results show that, for the proposed design, $\pi_k \approx \lambda_k$, see Figure 1. Furthermore, the simulation shows that the relative empirical bias, see Table 2, is relatively small for all three estimators; in absolute value the largest is about 0.07 per cent. Hence, the estimators used for the proposed design and Pareto $\pi$ps design seem to work well with respect to empirical bias.

Figure 1: Monte Carlo estimated factual first-order inclusion probabilities for the 2P$\pi$ps design versus $\lambda_k$, population $U_1$, 30 000 replicates

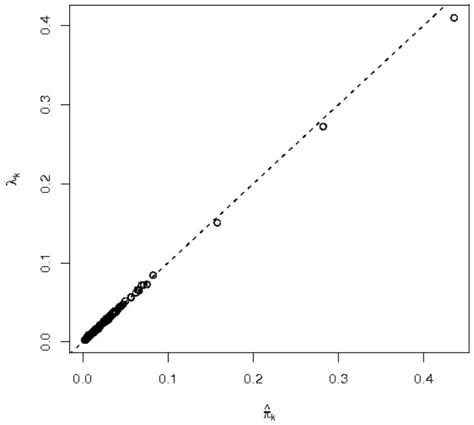

Table 2: Relative empirical bias for $\hat{t}_{PAR,\lambda}$, $\hat{t}_{2P\pi ps,\lambda}$ and $\hat{t}_{2P_{PO,SI},\pi^\star}$

| | | $\hat{t}_{PAR,\lambda}$ | $\hat{t}_{2P\pi ps,\lambda}$ | $\hat{t}_{2P_{PO,SI},\pi^\star}$ |
|---|---|---|---|---|
| $n = 5$ | $U_1$ | 0.000272 | -0.000728 | 0.000540 |
| | $U_4$ | -0.002537 | -0.001469 | 0.002128 |
| $n = 5000$ | $U_4$ | 0.000009 | -0.000001 | -0.000017 |

The simulation results also show that using any of the three designs coupled with the $\pi$-estimator is more efficient than using a SI design with the $\pi$-estimator, see Table 3. Furthermore, the two $\pi$ps designs are more efficient, in terms of precision, than the two-phase design although the discrepancies decreases as the overall sample fraction decreases. Finally, for all sample sizes and for all populations in the simulation, the proposed sampling scheme, and hence the proposed design with the reciprocal of (7) as design weights in (2), behave equally well, in terms of precision, as the Pareto $\pi$ps design.

An estimator of the variance of the $\pi$-estimator for the 2P$\pi$ps design has not yet been derived. However, one option is to use the variance estimator proposed for the Pareto $\pi$ps design. It seem to work well with respect to empirical bias as shown in Table 4.

Table 3: Empirical design effect of Pareto $\pi$ps, 2P$\pi$ps and 2P$_{PO,SI}\pi^\star$ compared to SI design

| | | $\hat{t}_{PAR,\lambda}$ | $\hat{t}_{2P\pi ps,\lambda}$ | $\hat{t}_{2P_{PO,SI},\pi^\star}$ |
|---|---|---|---|---|
| $n=5$ | $U_1$ | 0.000551 | 0.000545 | 0.028207 |
| | $U_4$ | 0.000009 | -0.000001 | -0.000017 |
| $n=5000$ | $U_4$ | 0.000547 | 0.000559 | 0.028413 |

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

| | | $\hat{t}_{PAR,\lambda}$ | $\hat{t}_{2P\pi ps,\lambda}$ | $\hat{t}_{2P_{PO,SI},\pi^\star}$ |
|---|---|---|---|---|
| $n=5$ | $U_1$ | 0.008119 | -0.003083 | 0.023153 |
| | $U_4$ | -0.002537 | -0.001469 | 0.002128 |
| $n=5000$ | $U_4$ | 0.008022 | -0.013745 | -0.002023 |

# 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, which is more efficient, in terms of precision, than using a SI design coupled with the $\pi$-estimator.

The proposed scheme can be used as an approximate method to generate a $\pi$ps sample. Simulation result show that using the proposed design coupled with the $\pi$-estimator 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. However, the magnitude of the design effect is dependent on the choice of size variable. In the simulation at hand, the co-variation between the study variable and the auxiliary information is high; in $U_1$ $\rho(y,x) = 0.998$. In cases when the co-variation is smaller, less gain, in terms of precision, compared to using a SI design would be expected.

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 used are implemented in most statistical softwares. For the future the properties of the inclusion probabilities for the proposed design need to be further studied, particular in order to make a design-based variance estimation feasible.

## References

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

Hájek, J. (1964) Asymptotic Theory of Rejective Sampling with Varying Probabilities From a Finite Population. *The Annals of Mathematical Statistics.*

Kröger, H., Särndal, C.-E. & Teikari, I. (2003) Poisson Mixture Sampling Combined with Order Sampling. *Journal of Official Statistics*, **19**1, 59-70.

Neyman, J. (1938) Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association*, **33**, 101-116.

Rosén, B. (1997a) Asymptotic Theory for Order Sampling. *Journal of Statistical Planning and Inference*, **62**, 135-158.

Rosén, B. (1997b) On Sampling with Probability Proportional to Size. *Journal of Statistical Planning and Inference*, **62**, 159-191.

Rosén, B. (2000) On Inclusion Probabilities for Order $\pi$ps Sampling. *Journal of Statistical Planning and Inference*, **90**, 117-143.

Saavedra, P. J. (1995) *Fixed Sample Size pps Approximations with Permanent Random Number.* 1995 Joint Statistical Meetings, American Statistical Association, Orlando, Florida, USA. **62**, 159-191.

Sampford, M. R. (1967) On Sampling Without Replacement with Unequal Probabilities of Selection. *Biometrika*, **54**, 499-513.

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