Probability-Proportional-to-Size Sampling from a Rare Population

Jens Olofsson*

Abstract
This paper presents an alternative to existing procedures used when sampling from rare populations. It is based on a two-phase sampling scheme with a corresponding probability-proportional-to-size sampling design. The sampling design proposed has the frequently used conditional Poisson sampling design as a special case. The proposed procedure is applied to a real life survey situation where real estates with fishing rights in Sweden constitute the population of interest.

Key Words: Two-phase sampling scheme, probability-proportional-to-size sampling design, rare populations, estimation

1. Introduction

In more and more situations information on different rare populations is needed. This is a challenge for the survey statistician since the data source(s) available to be used as sampling frame(s) does not information on the characteristics which define such a rare population. Kalton (2009) discuss different ways of handling this in a recent overview.

An all to common solution in practice is to use a larger sampling fraction than what would have been used in case of full information, and combine this with some kind of initial screening procedure in order to obtain survey results with reasonable precision.

Such a solution increases the overall survey cost as well as the response burden compared to the situation where the information on those characteristics defining the rare population of interest is complete in what is used as a sampling frame.

In this paper an alternative procedure is presented, which is based on a two-phase approach with a corresponding fixed-size probability-proportional-to-size sampling design. The procedure is applied to a real life survey situation where real estates with fishing rights as disposal rights in Sweden constitute the rare population of interest.

2. Probability-proportional-to-size sampling

Several methods on how to generate a fixed-size πps sample as been proposed in the statistical literature. See Brewer and Hanif (1983) for an overview and Tillé (2006) for more on both old and more recent πps designs.

In practice strict πps designs have rarely been used, due to difficulties with the implementation, e.g. Sampford (1967). Instead approximative πps designs as the Conditional Poisson Sampling (CPS) design proposed by Hájek (1964) or the Pareto πps sampling (PAR) design proposed by Rosén (1997a, 1997b).

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

*Department of Statistics, Örebro University, Fakultetsgatan 1, 701 82 Örebro, Sweden

design attain the target inclusion probabilities of a strict fixed-size πps sample asymptotically.

Olofsson (2010a) generalized the design and derived algorithms for calculating first- and second-order inclusion probabilities of the 2Pπps design efficiently.

2.1 The 2Pπps sampling design

Neyman (1938) introduced the two-phase (2P), or double (DBL), sampling design as a way of gathering information in the first phase necessary for a stratification in the second phase. General formulas for variances and variance estimators, irrespective of sampling designs in each phase, were derived by Särndal and 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 and Hurwitz (1946). See also Särndal, Swensson, and 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 Poisson (PO) design with $\Pr(k \in S_0) = \lambda_{ak}$, such that $\sum_{k=1}^{N} \lambda_{ak} = m > 0$ and $\lambda_{ak} \propto x_k$.

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.

If $\{s_0^i\}_{i=1}^{\infty}$ be an infinite sequence of independent initial samples using a PO 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 $\tau = \min(i : n \leq |s_0^i| \leq M)$.

Note that a sufficient condition for eventually reaching the third step of the scheme is that $\Pr(n \leq |S_0| \leq M) > 0$.

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 $\Omega_s = \{s_a : s \subseteq s_a \subseteq U, |s_a| \leq M\}$, then the corresponding design of the sampling scheme presented above, the 2Pπps design, can be expressed as

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

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

As Laitila and Olofsson (2010) and Olofsson (2010a) states, is suggested to use the proposed design to, in an easy way, generate a πps sample.

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

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 proposed by Hájek (1964).

The first- and second-order inclusion probabilities of the 2Pπps design are given by

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

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

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

where $n_{S_0}^{-k,l} = |S_0 \setminus \{k,l\}|$, are the first- and second-order inclusion probabilities, respectively, of the 2Pπps design. A formal derivation can be found in Olofsson (2010a).

2.1.1 An example

In order to illustrate how well the 2Pπps design work, even in the standard setting with $M = N$, an example is here given where a well known auxiliary vector from the literature, viz. one of the vectors in Sampford (1967). See Table 1. The maximum integer-valued expected sample size m is here equal to 5.

When $n = 2$ and x_k is small or high the first-order inclusion probabilities of the 2Pπps design are closer to the target probabilities compared to those of the CPS design. 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. Olofsson (2010a) shows that an upper bound of the bias resulting from using the reciprocal of the target probabilities, $\lambda_{ak} n/m$, compared to the first-order inclusion probabilities, is smaller for the 2Pπps design with $m = 5$ and $M = 10$, compared to the CPS design.

3. Application

Within the intra disciplinary research program *Adaptive management of fish and wildlife* financed by the Swedish Environment Protection Agency, a research group from the Swedish University of Agricultural Sciences started a project with the objective to survey fishing right owners in order to obtain knowledge on their objective(s) with owning a real property with fishing rights, their expectancies and demand for revenue, the usage of their fishing rights and so on.

However, there exists no collective source of information as a register or such of fishing right owners, but per definition they are owners of at least one real estate with fishing rights as a right of disposal.

Table 1: Population values and probabilities

k	y_k	x_k	$\lambda_k n/m$	π_k^{CPS}	$\pi_k^{2P\pi ps}$
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

Using the Swedish national land register as a sampling frame, and information within it a pilot survey was done in order to get estimates on the number of real estates in Sweden by the end of 2008 with fishing rights as rights of disposal and major domains as well as indicators of a real estate having fishing rights as rights of disposal.

The design was a disproportionate stratified sampling design, with 84 strata created based on the localization of the real estates (14 regions) and available auxiliary information in the register (6 groups). For the last group the only available information was the total area of the real estates. Hence, a fixed-size πps design as the 2Pπps design seemed to be an appropriate design to use within those strata belonging to this group of real estates. The other sampling designs used were simple random sampling (SI) design except for 14 strata which were take-all, or certainty, strata. More on the survey design can be found in Olofsson (2010b).

3.1 Estimation

In the presence of full response the unknown population total $\hat{t}_y$ could be estimated using the H-T estimator proposed by Horvitz and Thompson (1952) and discussed thoroughly in e.g. Särndal et al. (1992).

On the other hand, in presence of nonresponse, using the H-T estimator will give rise to nonresponse bias of unknown size of direction and magnitude as well as an over-estimation of the variance, since the variables of interest are only observed for the response set r.

If there exists auxiliary information ($\mathbf{x}_k$) it is possible to use the calibration method to adjust for the nonresponse. Deville and Särndal (1992) derived a general regression estimator calibrating the designs weights to known totals for the population. The calibration method was further developed in the context of nonresponse by Särndal and Lundström (2005) from which the notation here is adopted. The calibration method can utilize information on the population level ($\mathbf{x}_k^\star$) and/or on the sample level ($\mathbf{x}_k^\circ$). If auxiliary information is used at the population level the method requires that the vector of population totals $\sum_U \mathbf{x}_k^\star$ is known and that $\mathbf{x}_k^\star$ is known for every $k \in r$. On the other hand, if the auxiliary information is used at

the sample level it is required that $\mathbf{x}_k^\circ$ is known for every $k \in s$. In case of the information is used at both levels the auxiliary vector becomes $\mathbf{x}_k = \binom{\mathbf{x}_k^\star}{\mathbf{x}_k^\circ}$ of dimension $G^\star + G^\circ$.

The idea is to find a set of weights w_k such that $\sum_{h=1}^H \sum_{r_h} w_k \mathbf{x}_k = \mathbf{X} = \binom{\mathbf{X}^\star}{\hat{\mathbf{X}}^\circ}$. In order to utilize information on both levels Särndal and Lundström (2005) suggest three different estimators using the same input information $\binom{\mathbf{X}^\star}{\hat{\mathbf{X}}^\circ}$ viz. the single-step procedure, the two-step A procedure and the two-step B procedure.

In the two-step B procedure the intermediate weights w_k° are computed by calibration from r to s such that $\sum_{h=1}^H \sum_{r_h} w_k^\circ \mathbf{x}_k^\circ = \sum_{h=1}^H \sum_{s_h} d_k \mathbf{x}_k^\circ = \hat{\mathbf{X}}^\circ$, where $w_k^\circ = v_k^\circ d_k$ and

$$\begin{aligned} v_k^\circ &= 1 + \left(\sum_{h=1}^H \sum_{s_h} d_k \mathbf{x}_k^\circ - \sum_{h=1}^H \sum_{r_h} d_k \mathbf{x}_k^\circ \right)' \times \\ &\quad \times \left(\sum_{h=1}^H \sum_{r_h} d_k \mathbf{x}_k^\circ \left(\mathbf{x}_k^\circ\right)' \right)^{-1} \mathbf{x}_k^\circ \end{aligned} \tag{1}$$

in accordance with expression (11.16) in Särndal and Lundström (2005).

In the second step the weights given by (1) are used as initial weights. The final weights are computed by a calibration from s to U such that $\sum_{h=1}^H \sum_{r_h} w_k \mathbf{x}_k^\star = \sum_{h=1}^H \sum_{r_h} v_k v_k^\circ d_k \mathbf{x}_k^\star = \mathbf{X}^\star$, where

$$\begin{aligned} v_k &= 1 + \left(\sum_{h=1}^H \sum_{U_h} \mathbf{x}_k^\star - \sum_{h=1}^H \sum_{r} w_k^\circ \mathbf{x}_k^\star \right)' \times \\ &\quad \times \left(\sum_{h=1}^H \sum_{r_h} w_k^\circ \mathbf{x}_k^\star \left(\mathbf{x}_k^\star\right)' \right)^{-1} \mathbf{x}_k^\star \end{aligned} \tag{2}$$

which is in accordance with expression (11.19) in Särndal and Lundström (2005).

An estimator of the unknown population total t_y is given by

$$\hat{t}_{yw2B} = \sum_{h=1}^H \hat{t}_{hw2B},$$

where

$$\hat{t}_{hw2B} = \sum_{r_h} w_k y_k$$

and $w_k = v_k v_k^\circ d_k$, with v_k given by (2) and v_k° by (1). A corresponding variance estimator is given by

$$\widehat{var}(\hat{t}_{yw2B}) = \sum_{h=1}^H \widehat{var}_{1h} + \widehat{var}_{2h}, \tag{3}$$

where the sampling component $\widehat{var}_{1h}$ is given by

$$\begin{aligned} \widehat{var}_{1h} &= \sum\sum_{r_h} (d_k d_l - d_{kl})(v_k^\circ v_k \hat{e}_k)(v_l^\circ v_k \hat{e}_l) - \\ &\quad - \sum_r d_k (d_k - 1) v_k^\circ (v_k^\circ - 1)(v_k \hat{e}_k)^2 \end{aligned} \tag{4}$$

and the nonresponse component $\widehat{var}_{2h}$ by

$$\widehat{var}_{2h} = \sum_r v_k^\circ (v_k^\circ - 1)(d_k v_k \hat{e}_k)^2 \tag{5}$$

with

$$\hat{e}_k = y_k - \left(\mathbf{x}_k^\star\right)' \left(\sum_r d_k v_k^\circ \mathbf{x}_k^\star \left(\mathbf{x}_k^\star\right)' \right)^{-1} \sum_r d_k v_k^\circ \mathbf{x}_k^\star y_k$$

under the assumptions that the respondents response independently, i.e. $\Pr(k, l \in R|S = s) = \Pr(k \in R|S = s)\Pr(l \in R|S = s) = \theta_k\theta_l$ for all $k \neq l \in s$ and that $1/v_k^\circ$ can be used as a proxy variable for the unknown θ_k's in accordance with Särndal and Lundström (2005).

The estimator (4) of the sampling component in the variance of $\hat{t}_{hw2B}$ utilize the second-order inclusion probabilities obtained under the used design, which implies that the computation involves $\sum_{h=1}^{H} m_h(m_h - 1)/2$ terms. Although it is possible to calculate the second-order inclusion probabilities of the 2Pπps design as shown by Olofsson (2010a), it is computer intensive.

An alternative to using (4) as an estimator of the sampling component of (3) is to use the Hájek approximation, , as stated by Särndal and Lundström (2005), assuming $\sum_{s_h}(1 - \pi_k)/\sum_{U_h}\pi_k(1 - \pi_k) = 1$ for every $h = 1, 2, \ldots, H$. The estimator of the approximative variance, see Hájek (1964), is given by

$$\widehat{var}_J = \frac{n_h}{n_h - 1}\sum\nolimits_{r_h}\left(d_k v_k \hat{e}_k - \frac{\sum_{r_h} d_k v_k^\circ \hat{e}_k(1 - 1/d_k)}{\sum_{r_h}(1 - 1/d_k)v_k^\circ}\right)^2 (1 - 1/d_k)v_k^\circ. \quad (6)$$

Hence, an alternative variance estimator for $\hat{t}_{yw2B}$ is given by

$$\widehat{var}_{alt}(\hat{t}_{yw2B}) = \sum\nolimits_{h=1}^{H} \widehat{var}_J + \widehat{var}_{2h},$$

where $\widehat{var}_J$ is given by (6) and $\widehat{var}_{2h}$ by (5).

3.2 Results

A questionnaire was sent to the owner the 5 965 sampled real estates. If the owner was a physical person the questionnaire was sent by post and consisted of 31 items which were combined into six blocks. The questionnaire had three skip questions and the questions could be closed, open, or have an opened ending. To the owners of the remaining real estates a shortened questionnaire (seven items) was sent by post or electronically.

The questionnaire consisted of questions about the fishing rights, the amount of fish catched on the real estate and other kinds of activities on the real estate related to the fishing rights. The remaining block of questions were questions on the owner(s) objective with owning a real estate with fishing rights, the management and fishing right management associations as well as background information and general questions regarding the owner's attitude toward some statements on issues related to fishing rights.

In order to reduce the nonresponse, two remainder cards where sent out as well as a new questionnaire. In despite of the efforts, the study suffers from nonresponse. As mentioned earlier nonresponse causes nonresponse bias of unknown direction and magnitude as well as an over-estimation of the variance.

The nonresponse rate, weighted as well as unweighted, is a function of the response rate. The unweighted nonresponse rate is defined as $1 - r_u$, where

$$r_u = \frac{\sum_{h=1}^{H}\sum_{r_h} 1}{\sum_{h=1}^{H}\sum_{s_h} 1},$$

whilst the weighted is defined as $1 - \tilde{r}_u$, where

$$\tilde{r}_u = \frac{\sum_{h=1}^{H}\sum_{r_h} 1/d_k}{\sum_{h=1}^{H}\sum_{s_h} 1/d_k}.$$

For the survey presented here the unweighted nonresponse rate was 47 percent compared to 33 percent if weighted. The reason for the former being larger than the latter is due to the fact that real estates selected with small inclusion probability, i.e. large design weight, responded to a lesser extent.

The item response was handled by means of imputation, since full information on the response set r is a requirement in order to be able to use the calibration method as a tool of adjustment for nonresponse as presented and discussed by Särndal and Lundström (2005).

The main variable was if the real estates had fishing rights as a right of disposal or not. Based on the three principles listed by Särndal and Lundström (2005) an auxiliary vector consisting of information on population as well as sample level were chosen.

At the population level the regional location of the real estates was used since it defines one of the major domains of interest, although the nonresponse analysis indicates that some regions could be collapsed. However maintaining all the regions seemed to satisfy the second principle stated by Särndal and Lundström (2005). Hence, no collapsing was done.

At the sample level the regional location was also used, as well as three different dichotomous variables; auxpriv, auxsvea and auxarea. The first assumes value 1 if the real estate is privately owned and 0 otherwise. The second assumes value 1 if the owner is Sveaskog AB and 0, whilst the third assumes value 1 if any kind of area except for water and total area is known for the real estate and 0 otherwise. All three of the variables seemed to comply with the three principles listed by Särndal and Lundström (2005).

By using the calibration method to adjust for the nonresponse and the two-step B procedure, the total number of real estates with fishing rights as right of disposal in Sweden by 2008 was estimated to be 404 751 with an standard error of 41 016. The estimated coefficient of variation equal to 10.1 percent indicates that the estimate has an acceptable precision.

The point estimate of the proportion was 14.3 percent (standard error 1.5), which imply that the population of interest here does not qualify to be a rare population according to the definition by Kish (1987). However it is about seven percentage units less than the unweighted proportion. The interpretation is that real estates with fishing rights has been selected with an higher probability than the real estates without fishing rights. This pattern is emphasized looking only at the sixth group of real estates, for which only the total area of the real estate was known and the 2Pπps design used as sampling design. There the proportions were 24 and 18 percent, respectively. Hence, this indicates that the design of the survey has worked as intended, and that a fixed-size πps design as the 2Pπps design was a good choice for the problem at hand.

4. Discussion

This paper presents a fixed-size probability-proportional-to-size sampling design, the 2Pπps design. The design has comparable theoretical properties to those fixed-size πps designs commonly used in practice and has the CPS design as a special case.

It has been showed that if there exists an auxiliary variable in shape of some size variable, the 2Pπps design is possible to apply in a real life survey situation, due to its easy implementation, in order to obtain estimates on what was thought

to be a rare population with good result.

The present study yields the first estimates on the number of real estates with fishing rights as rights of disposal in Sweden.

References

Brewer, K., & Hanif, M. (1983). *Sampling with unequal inclusion probabilities* (Vol. 15). New York: Springer-Verlag.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Hájek, J. (1964, December). 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.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. *Survey Methodology*, *35*(2), 125–141.

Kish, L. (1987). *Statistical design for research.* New York: John Wiley & Sons, Inc.

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.

Olofsson, J. (2010a). *Algorithms to find exact inclusion probabilities for the $2p\pi ps$ sampling designs.* (manuscript)

Olofsson, J. (2010b). *A survey design to find and estimate the number of real estates with fishing rights in sweden.* (manuscript)

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.

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

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse.* Chichester, England: Wiley.

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.* New York: Springer Verlag.

Tillé, Y. (2006). *Sampling algorithms.* Springer.