The First Phase Order Sampling for the Second Phase Stratification

Danutė Krapavickaitė
Vilnius Gediminas Technical University, Vilnius, Lithuania, LT 10223
e-mail: danute.krapavickaite@vgtu.lt

Abstract

A two-phase sampling design with order sampling in the first phase and stratified simple random sampling in the second phase is studied in the paper. Order sampling is a sampling design without replacement and with the elements selected with probabilities proportional to their size. The second-phase stratification using suitably chosen sample sizes in the strata allows us to equalize the design weights for elements. Calculation of the first-phase inclusion probabilities is computationally complicated, therefore the quasi-Horvitz-Thompson estimator of the total is used. Its approximate variance, the estimator of this variance and simulation results are presented.

Keywords: estimator of the total, estimator of variance, selection of households through individuals, successive sampling.

1. Introduction

An unequal probability sampling design allows us to take into account individual features of population elements assigning inclusion probabilities to the elements. This property is useful when it is possible to construct a population parameter estimator which together with the sample design yields an effective sampling strategy.

Sometimes unequal probability sampling arises naturally. It is the case of sampling households through individuals. Suppose that there is no list of households in the country, and the sampling frame of individuals contains their addresses. Sampling is carried out as follows: (i) individual is selected from the sampling frame with equal probabilities, and his/her household is included into the sample; (ii) the next individual is selected from the sampling frame with equal probabilities and independently from previous one. His/her household is included into the sample if it is not there. Otherwise this sampling step is ignored; (iii) the second step is repeated independently from the previous steps until the predetermined n size sample of households is selected. A household selection probability is proportional to the number of its individuals included into the sampling frame.

This sampling design in a general context is described by Rosén B. (1972) and called a successive sampling design. It is a fixed size sampling scheme without replacement, and the elements are selected independently with probabilities proportional to their size.

2. Successive sampling design

Before proceeding to the main results, let us present the successive sampling in more detail.

Let $\mathcal{U} = \{1, 2, ..., N\}$ be a finite population of unknown size N, and let m be the size measure of elements with the values m_i, the sum of which $m_1 + \ldots + m_N = M$ is known. Let us denote by $p_i = m_i/M$ the normalized element size measure, $i = 1, ..., N$. Let s be the element sample of size n, which, according to Rosén (1972), is selected as follows:

(i) The elements are drawn from $\mathcal{U}$ successively and with replacement until n different elements have been selected, and these elements constitute the element-sample,

(ii) At each draw, the element i is selected with probability p_i, and all the draws are made independently of one another.

Rosén (1972) has called this sampling design successive and obtained approximation formula of the the element inclusion probability for a successive sample:

$$\pi_i = \pi_i(n) = 1 - \mathrm{e}^{-p_i n} + \frac{r_i(n)}{N}, \quad r_i(n) \le r(n),\ i \in \mathcal{U}, \tag{1}$$

with some function $r(n)$.

Let y be a study variable defined on a finite population with the values $y_1, y_2, ..., y_N$, and $\tau(y) = \sum_{i=1}^{N} y_i$ be their population total.

Assumptions for population and successive sample:

A) The sample size n is large and the population size is larger: $0 < n/N < 1$,

B) None of the probabilities p_i, $i \in \mathcal{U}$, is an outlier,

C) None of the values $y_i - \bar{y}$, $i \in \mathcal{U}$, is an outlier as compared with the majority of deviations, $\bar{y} = \sum_{i=1}^{N} y_i/N$.

Theorem of Rosén (1972) asserts that if there is a sequence of populations $\mathcal{U}_k$ and samples $\mathbf{i}_k$, $\mathbf{i}_k \subset \mathcal{U}_k$, drawn under successive sampling design, of sizes $N_k \to \infty$, $n_k \to \infty$ for $k \to \infty$, which satisfy the conditions A)-C), and sizes of those finite populations and samples satisfy the inequality

$$0 < \liminf_{k\to\infty} \frac{n_k}{N_k} \le \limsup_{k\to\infty} \frac{n_k}{N_k} < 1,$$

then the sample sum $Z_{n_k} = \sum_{i\in\mathbf{i}_k} y_i$ is asymptotically normally distributed, its mean and variance are given in the paper of Rosén, and it is shown that

$$\lim_{k\to\infty} \max\ r(n_k) = 0.$$

Therefore, it follows from (1) that $\pi_i(n) \approx 1 - \mathrm{e}^{-p_i n} \approx np_i$. We see that inclusion probabilities $\pi_i(n)$ for successive sampling are only approximately equal to $\lambda_i(n) = np_i$, called target inclusion probabilities.

3. A class of order sampling designs

Successive sampling is a case of a wider class of sampling designs called order sampling designs, introduced also by Rosén (1996). Let $F_1, ..., F_N$ be absolutely continuous ordering distribution functions increasing in the interval $[0, \infty)$ with the densities $f_1, ..., f_N$. Sampling is carried out as follows.

(i) Independent random variables $Q_1, ..., Q_N$ are chosen with distribution functions $F_1, ..., F_N$, and these functions $F_1, ..., F_N$ are associated with the population elements $\mathcal{U} = \{1, ..., N\}$, respectively.

(ii) The variables $Q_1, ..., Q_N$ are realized and population elements with the smallest values Q constitute the sample.

We will study the fixed shape order distribution functions $F_i(t) = H\big(tH^{-1}(\lambda_i)\big)$, $i = 1, 2, ..., N$, with a given absolutely continuous shape distribution function $H(t)$ increasing in the interval $[0, \infty)$, and density $h(t)$ such that $f_i(t) = h\big(tH^{-1}(\lambda_i)\big)H^{-1}(\lambda_i)$, where H^{-1} denotes the inverse function.

There are several order sampling designs that belong to this class:

The sequential Poisson sampling design with a uniform shape distribution function $H(t) = t,\ h(t) = 1$ for $t \in [0, 1)$ and $H(t) = 1,\ h(t) = 0$ for $t \in [1, \infty)$, $Q_i = H^{-1}(U_i)/H^{-1}(\lambda_i) = U_i/\lambda_i$, where U_i, $i = 1, 2, ..., N$, are random variables distributed uniformly in the interval $[0, 1]$.

The successive sampling design with an exponential shape distribution function $H(t) = 1 - e^{-t},\ h(t) = e^{-t}$ for $t \in [0, \infty)$, $Q_i = \ln(1 - U_i)/\ln(1 - \lambda_i)$.

The Pareto sampling design with a Pareto shape distribution function $H(t) = t/(1 + t),\ h(t) = 1/(1 + t)^2$ for $t \in [0, \infty)$, $Q_i = \big(U_i(1 - \lambda_i)\big)/\big(\lambda_i(1 - U_i)\big)$.

The problem to obtain exact inclusion probabilities for the order sampling design is hardly solvable, and $\pi_k(n) \neq \lambda_k(n)$ as were for the successive design. Rosén (2000) has shown that $\pi_k(n)/\lambda_k(n) \to 1$ as $n \to \infty$ for $k = 1, 2, ..., N$.

Expressions of inclusion probabilities $\pi_k(n)$ are given by Rosén (2000) for successive sampling. Aires (1999) presented recursive expressions of inclusion probabilities for any order sampling design. Expressions of inclusion probabilities for a sequential Poisson and Pareto sampling design are also presented in Bondesson et al. (2006). Unfortunately, all of them are computationally cumbersome and can be calculated only for small populations. Therefore Horvitz-Thompson estimator $\widehat{\tau}(y) = \sum_{i \in \mathbf{i}} y_i/\pi_i$ of the total $\tau(y) = \sum_{i=1}^{N} y_i$ in the case of order sample $\mathbf{i}$ cannot be used in practical situations.

Let us consider another estimator of population total:

$$\widehat{\tau}(y)_\lambda = \sum_{i \in \mathbf{i}} \frac{y_i}{\lambda_i}. \tag{2}$$

called quasi-Horvitz-Thompson.

Relying on some limit theorem, Rosén has formulated a proposition giving an approximate expression for variance of estimator (2) and suggested the estimator of variance.

Approximation result (Rosén, 1996) for distribution of the quasi-Horvitz-Thompson estimator of total in the case of order sampling with a fixed distribution shape function. *Under conditions* A)-C)

(i) *The distribution of* $\widehat{\tau}(y)_\lambda$ *is well approximated by a normal distribution* $\mathcal{N}\big(\tau(y), \sigma^2(y; \lambda; H)\big)$ *with*

$$Var(\widehat{\tau}(y)_\lambda) \approx \sigma^2(y; \lambda; H) = \frac{N}{N-1} \sum_{i=1}^{N} \left(\frac{y_i}{\lambda_i} - \sum_{j=1}^{N} \frac{y_j}{\lambda_j} \frac{\alpha_j}{A} \right)^2 \lambda_i(1 - \lambda_i), \tag{3}$$

$$\alpha_j = \alpha_j(\lambda; H) = h\big(H^{-1}(\lambda_j)\big)H^{-1}(\lambda_j), \quad j = 1, ..., N, \quad A = \sum_{k=1}^{N} \alpha_k.$$

(ii) $\widehat{\tau}(y)_\lambda$ *is a consistent estimator for* $\tau(y)$.
(iii) $Var(\widehat{\tau}(y)_\lambda)$ *is well approximated by* $\sigma^2(y;\lambda;H)$.
(iv) *A good variance estimator for* $\sigma^2(y;\lambda;H)$ *is provided by*

$$\widehat{Var}(\widehat{\tau}(y)_\lambda) \approx \widehat{\sigma}^2(y;\lambda;H) = \frac{n}{n-1}\sum_{i\in\mathbf{i}}\Big(\frac{y_i}{\lambda_i} - \sum_{j\in\mathbf{i}}\frac{y_j}{\lambda_j}\frac{a_j}{\sum_{k\in\mathbf{i}}a_k}\Big)^2(1-\lambda_i), \qquad (4)$$

$$a_j = a_j(\lambda_j;H) = \alpha_j(\lambda_j;H)/\lambda_j, \quad j = 1,2,...,N.$$

The weights of the quasi-Horvitz-Thompson estimator are not equal, and sometimes it may be not desirable, especially when the unequal probability sampling design arises naturally as in the case of the successive sampling design.

4. Two phase sampling design

In order to equalize the weights in the quasi-Horvitz-Thompson estimator of total (2), the second phase simple random stratified sampling design with probabilities conversely proportional to the household size is used. This kind of second phase stratification has been used by Ilves (2005) for the Horvitz-Thompson estimator of the total.

We describe the second phase sampling design. Let $\mathbf{i}^{(1)}$ denote the 1st phase order sample with a fixed shape ordering distribution of size $n^{(1)}$. Let us divide this sample into strata: $\mathbf{i}^{(1)} = \mathbf{i}_1^{(1)} \cup ... \cup \mathbf{i}_G^{(1)}$, so that $m_i = g$ for $i \in \mathbf{i}_g^{(1)}$, $n_g^{(1)}$ is the size of $\mathbf{i}_g^{(1)}$, and $n^{(1)} = n_1^{(1)} + ... + n_G^{(1)}$. A simple random stratified sample $\mathbf{i}^{(2)} = \mathbf{i}_1^{(2)} \cup ... \cup \mathbf{i}_G^{(2)}$, $\mathbf{i}^{(2)} \subset \mathbf{i}^{(1)}$ with $\mathbf{i}_g^{(2)} \subset \mathbf{i}_g^{(1)}$ of size $n_g^{(2)} = \Big[n_g^{(1)}/g\Big]$, $g = 1,....,G$, is used in the second phase.

Quasi-Horvitz-Thompson estimator (2) may be rewritten:

$$\widehat{\tau}(y)_\lambda = M\sum_{g=1}^{G}\frac{n_g^{(1)}}{n^{(1)}g}\bar{y}_g^{(1)}. \qquad (5)$$

Here $\bar{y}_g^{(1)} = \sum_{j\in\mathbf{i}_g^{(1)}} y_j/n_g^{(1)}$, is the mean of the 1st phase poststratum, $g = 1,2,...,G$. We estimate it by $\bar{y}_g^{(2)} = \sum_{j\in\mathbf{i}_g^{(2)}} y_j/n_g^{(2)}$, and the estimator, derived from (5) for a two-phase sampling design is

$$\widehat{\widehat{\tau}}(y) = M\sum_{g=1}^{G}\frac{n_g^{(1)}}{n^{(1)}g}\bar{y}_g^{(2)}. \qquad (6)$$

It is not unbiased, but only an approximately unbiased estimator of the population total $\tau(y)$. Its weights are approximately equal:

$$w_i^{(2)} = \frac{1}{\lambda_i}\frac{n_g^{(1)}}{n_g^{(2)}} \approx \frac{M}{n^{(1)}}, \quad \text{for } i \in \mathbf{i}_g^{(2)}, \quad g = 1,...,G.$$

We are ready to formulate our result now.

Proposition. *Suppose that the sampling design is with the first phase fixed size order sampling having the fixed order distribution shape, and the second phase sampling design is stratification by the integer-valued size measure* $g \in \{1,2,...,G\}$ *with*

the sampling level conversely proportional to the element size and simple random sampling in the strata. The approximate variance of quasi-Horvitz-Thompson estimator (6) *of the population total is derived:*

$$AVar\big(\widehat{\widehat{\tau}}(y)\big) = \sigma^2(y;\lambda;H) + M^2\sum_{g=1}^{G}\frac{1}{g^2}E\left(\Big(\frac{n_g^{(1)}}{n^{(1)}}\Big)^2\Big(1-\frac{n_g^{(2)}}{n_g^{(1)}}\Big)\frac{s_g^{(1)2}}{n_g^{(2)}}\right), \qquad (7)$$

$$\begin{aligned}\sigma^2(y;\lambda,H) &= \frac{N}{N-1}\left(\frac{M}{n^{(1)}}\sum_{g=1}^{G}(N_g-1)\frac{s_g^2}{g}(1-\lambda_{(g)})\right.\\ &+ \left.\frac{M}{n^{(1)}}\sum_{g=1}^{G}gN_g\Big(\frac{\bar{y}_g}{g}-\frac{1}{A}\sum_{l=1}^{G}\sum_{j\in\mathcal{U}_l}\frac{y_j\alpha_j}{l}\Big)^2(1-\lambda_{(g)})\right),\end{aligned}$$

$$s_g^2 = \frac{1}{N_g-1}\sum_{j\in\mathcal{U}_g}(y_j-\bar{y}_g)^2, \quad \bar{y}_g = \frac{1}{N_g}\sum_{j\in\mathcal{U}_g}y_j, \quad \lambda_{(g)} = \frac{n^{(1)}g}{M}, \quad g=1,2,...,G.$$

We propose the following estimator for variance (7) *of the estimator* (6):

$$\begin{aligned}\widehat{Var}(\widehat{\widehat{\tau}}(y)) &= \frac{M^2}{n^{(1)}-1}\sum_{g=1}^{G}\frac{1}{g^2}\Big(\frac{n_g^{(1)}}{n^{(1)}}-\frac{g}{M}\Big)(1-\lambda_{(g)})\widehat{s}_g^{(2)2}\\ &+ \frac{M^2}{n^{(1)}-1}\sum_{g=1}^{G}\frac{n_g^{(1)}}{n^{(1)}}\Big(\frac{\bar{y}_g^{(2)}}{g}-\widehat{R}\Big)^2(1-\lambda_{(g)})\\ &+ M^2\sum_{g=1}^{G}\frac{1}{g^2}\left(\Big(\frac{n_g^{(1)}}{n^{(1)}}\Big)^2\Big(1-\frac{n_g^{(2)}}{n_g^{(1)}}\Big)\frac{\widehat{s}_g^{(2)2}}{n_g^{(2)}}\right), \qquad (8)\end{aligned}$$

$$\widehat{s}_g^{(2)2} = \frac{1}{n_g^{(2)}-1}\sum_{j\in\mathbf{i}_g^{(2)}}(y_j-\bar{y}_g^{(2)})^2, \quad \bar{y}_g^{(2)} = \frac{1}{n_g^{(2)}}\sum_{j\in\mathbf{i}_g^{(2)}}y_j,$$

$$\widehat{R} = \frac{1}{\sum_{l=1}^{G}n_l^{(1)}a_{(l)}/l}\sum_{l=1}^{G}\frac{a_{(l)}}{l^2}n_l^{(1)}\bar{y}_l^{(2)}, \quad a_{(l)} = \frac{\alpha(\lambda_{(l)},H)}{\lambda_{(l)}}.$$

The approximate variance is derived using conditional and unconditional variances and expectations. The middle term of the variance estimator is slightly biased.

5. Simulation results and conclusions

Real Labor Force survey data of Statistics Lithuania consisting of 3009 households of size from 1 to 7 have been used for a simulation study as a survey population. There are 7836 individuals, 3186 employed and 484 unemployed, in the population. The average household size is 2.6. There are 2 study variables y: the number of employed and the number of unemployed individuals in the household with the values $y_i \in \{0,1,...,m_i\}$. The correlation coefficients of these variables with the household size are 0.58 and 0.18, respectively. The estimates of the totals are studied. The samples have been selected according to 3 sampling designs:

(i) A simple random sample (SI) of households of $n = 300$ size. The Horvitz-Thompson estimator is used here.

(ii) The household sample of $n = 300$ size under the order sampling designs (denoted by Succ., Pareto, Poisson). The total $\tau(y)$ is estimated by $\widehat{\tau}(y)_\lambda$, its variance $Var\big(\widehat{\tau}(y)_\lambda\big)$ is estimated by (4).

(iii) Two-phase sampling design of $n = 790$ size with order household sample design in the first phase and a stratified simple random sampling design with a sampling level, inverse to the household population size, in the second phase (denoted by Succ. 2ph, Pareto 2ph, Poisson 2ph). The first phase sample size is chosen so that the second phase sample size could be close to 300 (sample size in the one phase sampling design). (6) and (8) estimators are used for estimation.

The selection of samples has been repeated for $B = 10000$ times, the number of employed and unemployed individuals in the population has been estimated. For $\widehat{\theta} = \widehat{\tau}(y)_\lambda$, $\overline{\widehat{\tau}}(y)$ average of estimates of the total $\overline{\widehat{\theta}}$, true or approximate variance $Var(\widehat{\theta})$, the average $\overline{\widehat{Var}}(\widehat{\theta})$ of the estimates of variances and empirical variance $Var_{em}(\widehat{\theta})$ is presented in Table 1 for all sample designs for employed and unemployed individuals.

Table 1. Estimates of totals and variances

	Employed				Unemployed			
Sampling design	$\overline{\widehat{\theta}}$	$Var(\widehat{\theta})$	$\overline{\widehat{Var}}(\widehat{\theta})$	$Var_{em}(\widehat{\theta})$	$\overline{\widehat{\theta}}$	$Var(\widehat{\theta})$	$\overline{\widehat{Var}}(\widehat{\theta})$	$Var_{em}(\widehat{\theta})$
SI	3185	25749	25751	24997	484	5059	5067	5084
Succ.	3186	17656	17644	17704	484	4749	4746	4914
Succ. 2ph	3188	–	16309	16632	484	–	4839	4958
Pareto	3186	17656	17646	17736	485	4749	4758	4784
Pareto 2ph	3184	–	16273	15811	483	–	4839	4960
Poisson	3185	17654	17653	17579	484	4749	4746	4670
Poisson 2ph	3184	–	16295	16673	484	–	4847	4950

Simulation results show that variance estimates are close to variances and empirical variances. The behavior of estimates, their approximate variances and the variance estimates is similar for all order sampling designs studied. Two-phase sampling design is efficient to estimate number of employed individuals and inefficient to estimate number of unemployed individuals because of high correlation between the number of employed individuals in the household and the household size, and low correlation between the number of unemployed individuals in the household and the household size.

References

Aires, N. (1999) Algorithms to Find Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs. *Methodology and Computing in Applied Probability*, 1:4, 457-469.

Bondesson, L., Traat, I. and Lundqvist, A. (2006) Pareto Sampling versus Sampford and Conditional Poisson Sampling. *Scandinavian Journal of Statistics*, 33, 699-720.

Ilves, M. (2005) Variance and its Estimator for a Practical Self-Weighting Two-Phase Design. In: *CD of Abstracts of the 55th Session of the International Statistical Institute (ISI), 5-12 April 2005, Sydney, Australia*, ISBN: 1877040282.

Rosén, B. (1972) Asymptotic Theory for Successive Sampling with Varying Probabilities Without Replacement, I, II. *The Annals of Math. Stat.*, 42(2), 373-397, 748-776.

Rosén, B. (1996) *On Sampling with Probability Proportional to Size.* Statistics Sweden: R&D Report, 1996:1.

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