Inclusion probabilities for successive sampling

Tomas Rudys[1]

[1]Statistics Lithuania, e-mail: tomas.rudys@gmail.com

Abstract

We give a short overview of calculation of first and second-order inclusion probabilities for successive sampling design. We compare the successive sampling first and second-order inclusion probabilities with already known numerical approximation results for Pareto πps and Conditional Poisson sampling designs. Pareto πps and successive sampling was introduced by Rosén and belongs to a class of sampling designs called order sampling with fixed distribution shape. At first an order sampling is introduced. We also give the examples of calculation of first and second-order inclusion probabilities for the mentioned above different sampling designs.

Keywords: Order sampling, successive sampling, first and second-order inclusion probabilities, numerical integration.

1 Introduction

Rosén (Rosén, 1996) studied and introduced a class of sampling designs called order sampling designs, which are executed as follows. Independent random variables, called ordering variables, are associated with the units in the population. A sample of size n is generated by first realizing the ordering variables, and then letting the units with the n smallest ordering variable values constitute the sample. Rosén also defined order sampling designs with fixed distribution shape: uniform, exponential, Pareto, successive. Author also derived the exact formulas for calculation of inclusion probabilities (Rosén, 1998) for these sampling designs.

Krapavickaitė (Krapavickaitė, 2012a) showed that Lithuanian Labour Force survey has successive sampling design and analysed the quality implementation actions for Lithuanian Labour Force Survey. Krapavickaitė also analysed order sampling designs and gave formulas for calculation of first and second-order inclusion probabilities for successive sampling (Krapavickaitė, 2012b).

Conditional Poisson and Pareto πps sampling designs were analysed and compared by Aires (1999) where the algorithms to find exact inclusion probabilities were derived. Author showed that it is feasible to calculate first and second-order inclusion probabilities for both sampling designs and program routines provide good numerical precision.

We compute first and second-order inclusion probabilities for successive sampling and compare them with first and second-order inclusion probabilities of Conditional Poisson and Pareto sampling designs. The successive sampling design was not studied very properly, maybe because it belongs to the same class of order sampling designs with fixed distribution shape as Pareto and Rosén showed that Pareto sampling design is optimal in the class of these sampling designs.

2 Order sampling

Consider a population $U = \{1, 2, ..., N\}$. For each unit i in the population is associated an independent random variable Q_i, called ranking variable, and a probability distribution function F_i, $[0, \infty)$, called order distribution, with density $f_i, i = 1, 2, ..., N$.

Order sampling from population U with sample size n, $n < N$, and order distributions $F_1, F_2, ..., F_N$ is carried as follows. Independent ranking variables $Q_1, Q_2, ..., Q_N$ with distributions $F_1, F_2, ..., F_N$ are realized. The units with the n smallest Q-values constitute the sample.
Let $H(t)$ be a probability distribution function with density $h(t) = H'(t), 0 \leq t < \infty$, and $\theta = (\theta_1, \theta_2, ..., \theta_N)$ are given real positive numbers – intensities. Together $H(t)$ and intensities θ denote the distribution functions $F_i, i = 1, ..., N$.
An order sampling design, $F_i, i = 1, ..., N$, is said to have fixed order distribution shape $H(t)$ with intensities θ, if following two equivalent conditions are met:

1. The ranking variables $Q_1, Q_2, ..., Q_N$ are of type $Q_i = Z_i/\theta_i, i = 1, ..., N$, where $Z_1, Z_2, ..., Z_N$ are independent, identically distributed (iid) random variables with common distribution $H(t)$.

2. The order distributions are $F_i(t) = H(\theta_i t)$, with density $f_i(t) = \theta_i h(\theta_i t)$, $0 \leq t < \infty, i = 1, ..., N$.

Denote $\lambda_1, \lambda_2, ..., \lambda_N$ as target inclusion probabilities for a, maybe approximate, πps sampling design with fixed sample size. Simply $\lambda_1, \lambda_2, ..., \lambda_N$ are given real numbers which satisfy: $0 < \lambda_i < 1, i = 1, ..., N, \sum_{i=1}^{N} \lambda_i = n$. It is shown by Rosén that using the order sampling design with fixed distribution shape, inclusion probabilities π_i can be approximately equal to given target inclusion probabilities $\lambda_i, i = 1, ..., N$.

3 First-order inclusion probabilities

Aires showed that first-order inclusion probabilities for different order sampling designs can be calculated using Lemma 2.
Lemma 2 (Aires, 1999, p. 461). Consider a sequence $Q_1, Q_2, ...$ of independent random variables with distribution functions $F_1, F_2, ...$ Let $Q_{(n)}^N$ be the n-th order statistic among $Q_1, Q_2, ..., Q_N$ with distribution function F_n^N. Then $F_n^N(t)$, $N = 1, 2, ..., n = 1, ..., N$, satisfy recursive equation:

$$F_n^N(t) = F_n^{N-1}(t) + F_N(t)\left(F_{n-1}^{N-1}(t) - F_n^{N-1}(t)\right), \tag{1}$$

where $F_0^N(t) = 1$, for all N and $t > 0$.
In the case of order sampling procedure, the probability of element N belonging to the sample s is:

$$\pi_N = P(N \in s) = P(Q_{(n)}^{N-1} > Q_N) = \int_0^\infty \left(1 - F_n^{N-1}(t)\right) f_N(t) dt. \tag{2}$$

The inclusion probability of any other unit i is derived similarly, from the corresponding formula for the rearranged sequence $Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ...,$
Q_N, Q_i instead.

3.1 The successive sampling case

Consider an order sampling design with fixed distribution shape. For successive sampling design the order distribution function can be expressed as $F_i(t) = H(\theta_i t) = 1 - e^{-\theta_i t}, \theta_i > 0$ for $i = 1, ..., N$. The densities then become $f_i(t) = F_i'(t) = (1 - e^{-\theta_i t})' = \theta_i e^{-\theta_i t}$. Parallel to this θ parametrization an alternative set or parameters which are more directly coupled to the inclusion probabilities was used (Aires, 1999): $\lambda_i = F_i(1) = 1 - e^{-\theta_i}, i = 1, ..., N$. This is motivated by the fact that λ_i approximates the inclusion probabilities in case $\sum_U \lambda_i = n$. Let $\tilde{\pi}_i$ denote the inclusion probabilities as functions of λ. Since the intensities, for successive sampling are $\theta_i = H^{-1}(\lambda_i) = -ln(1 - \lambda_i)$, then the probability of element N belonging to the sample s is:

$$\pi_N = -ln(1 - \lambda_N) \int_0^\infty \left(1 - F_n^{N-1}(t)\right) (1 - \lambda_N)^t dt. \tag{3}$$

The exact inclusion probabilities are computed according to Lemma 2, by numerical approximations with a computer program developed with statistical package R. The input of this program is a vector of given target inclusion probabilities $\lambda = (\lambda_1, \lambda_2, ..., \lambda_N)$. At first we compute F_n^N using the recursion in Lemma 2. For numerical integration we use adaptive Simpson$'$s and Monte-Carlo algorithms. The preprogrammed function for adaptive Simpson$'$s rule for numerical integration in statistical package R were used.
Example 1. The vector of target inclusion probabilities $\lambda = (0.1, 0.2, 0.3, 0.5, 0.9)$ is given. We compute first-order inclusion probabilities $\tilde{\pi_i}$ using successive sampling design. The population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^N \tilde{\pi_i} = 1.999999$, see Table 1.

3.2 The Pareto πps sampling case

Consider an order sampling design and suppose that $F_i(t) = H(\theta_i t) = \theta_i t/(1 + \theta_i t)$ is the standart Pareto distribution function with parameter $\theta_i > 0$ for $i = 1, ..., N$. Then the densities are $f_i(t) = \theta_i/(1 + \theta_i t)^2$. Since $\theta_i = H^{-1}(\lambda_i) = \lambda_i/(1 - \lambda_i)$, then the probability of element N belonging to the sample s is:

$$\pi_N = \lambda_N/(1 - \lambda_N) \int_0^\infty \left(1 - F_n^{N-1}(t)\right) \frac{1}{(1 + \lambda_N(t-1))^2} dt. \tag{4}$$

Example 2. We compute first-order inclusion probabilities for Pareto πps sampling design for the same target inclusion probabilities vector given in the example 1, with population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^N \tilde{\pi_i} = 1.999999$, see Table 1.

3.3 Conditional Poisson sampling case

Poisson sampling is a method for choosing a sample s of random size $|s|$, from a finite population U consisting of N elements. Each element i in the population has predetermined probability p_i of being included in the sample. A Poisson sample may be realised by using N independent Bernoulli trials to determine whether the element under consideration is to be included in the sample or not. Any experiment that results other that n out of the N elements being selected is rejected. One performs sequentially independent experiments until one of the experiments results in n out of N elements being selected.
First-order inclusion probabilities for conditional Poisson sampling can be calculated using Lemma 1 (Aires, 1999, p. 459).
Lemma 1. Consider a sequence of probabilities $p_1, p_2, ...$ and let $A_n(N)$ be the subset of all samples of size n among $\{1, ..., N\}$ for $n < N$. Then the quantities

$$S_n^N(p_1, ..., p_N) = \sum_{s \in A_n(N)} \prod_{i \in s} p_i \prod_{j \notin s} (1 - p_j)$$

with $N = 0, 1, 2, ...$ and $n = 0, ..., N$, may be calculated recursively by

$$S_n^N(p_1, ..., p_N) = p_N S_{n-1}^{N-1}(p_1, ..., p_{N-1}) + (1 - p_N) S_n^{N-1}(p_1, ..., p_{N-1})$$

for $n = 1, ..., N-1$ using the observations that $S_0^N = (1-p_1)(1-p_2)...(1-p_N)$ and $S_N^N = p_1 p_2 ... p_N$. The inclusion probability $\tilde{\pi_i}$ of any unt $i, i = 1, ..., N$, can be written as:

$$\tilde{\pi_i} = \frac{p_i S_{n-1}^{N-1}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}. \tag{5}$$

The first-order inclusion probabilities for conditional Poisson sampling design are calculated by a computer program developed with statistical package R. At first S_n^N are calculated using the recursion mentioned above. The input for the program is any vector of unconditional Bernoulli probabilities $p = (p_1, p_2, ..., p_N)$. As a result program returns conditional inclusion probabilities $(\tilde{\pi_1}, \tilde{\pi_2}, ..., \tilde{\pi_N})$.
Example 3. The vector of unconditional Bernoulli probabilities $p = (0.1, 0.2,$
$0.3, 0.5, 0.9)$ is given. We compute first-order conditional inclusion probabilities $\tilde{\pi_i}$, having population size $N = 5$ and sample size $n = 2$ elements. Notice that $\sum_{i=1}^N p_i = \sum_{i=1}^N \tilde{\pi_i} = 2$, see Table 1.

Table 1: First-order inclusion probabilities for different sampling designs

λ/p	$\tilde{\pi}_i$		
	Successive	Conditional Poisson	Pareto πps
0.1	0.087999779247	0.069470260223	0.094559623047
0.2	0.184249333826	0.154275092936	0.189740430046
0.3	0.290362518450	0.259990706319	0.289828419746
0.5	0.540796307599	0.573187732342	0.517952730787
0.9	0.896591938179	0.943076208178	0.907918436717
sum: 2.0	1.999999877303	2.000000000000	1.9999996403442

4 Second-order inclusion probabilities

Consider an order sampling design with population size N and sample size of n units. Then the bivariate inclusion probability of the units $N-1, N$ is given by:

$$\begin{aligned}\pi_{N-1,N} &= P(N-1 \in s, N \in s) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N)) = \\ &= \int_0^\infty \left(1 - F_{n-1}^{N-2}(t)\right) f_{max(Q_{N-1},Q_N)}(t)dt.\end{aligned} \tag{6}$$

Here

$$\begin{aligned}f_{max(Q_{N-1},Q_N)}(t) &= F'_{max(Q_{N-1},Q_N)}(t) = (F_{N-1}(t)F_N(t))' = \\ &= F'_{N-1}(t)F_N(t) + F_{N-1}(t)F'_N(t).\end{aligned}$$

The inclusion probability of an arbitrary pair of units $i < j$ may be determined by consideration of rearranged sequence $Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ..., Q_{j-1}$,
$Q_{j+1}, ..., Q_N, Q_i, Q_j$.

4.1 The successive sampling case

For calculation of second-order inclusion probabilities for successive sampling design we have the same order distribution functions and intensities notations as for the first-order inclusion probabilities. Then second-order inclusion probability for units $i < j$ can be expressed as follows:

$$\pi_{i,j} = \int_0^\infty \left(1 - F_{n-1}^{N-2}(t)\right) f_{max(Q_i,Q_j)}(t)dt, \tag{7}$$

where $f_{max(Q_i,Q_j)}(t) = \theta_i e^{-\theta_i t}(1 - e^{-\theta_j t}) + (1 - e^{-\theta_i t})\theta_j e^{-\theta_j t}$.
Example 4. We compute second-order inclusion probabilities $\tilde{\pi}_{i,j}$ for successive sampling design for the given target inclusion probabilities vector $\lambda = (0.1, 0.2,$
$0.3, 0.5, 0.9)$, where population size $N = 5$ and sample size $n = 2$ elements. The results are shown in Table 2. The control sum is $\sum_{i=1}^{N-1} \sum_{j=1}^{N} \tilde{\pi}_{i,j} = 1$.

4.2 The Pareto πps sampling case

Order distribution functions and intensities for calculation of second-order inclusion probabilities for Pareto sampling are the same as for the first-order ones. In this case second-order inclusion probability for units $i < j$ can be calculated using the equation 7. We give just the expression of the density function:

$$f_{max(Q_i,Q_j)}(t) = \frac{\theta_i}{(1+\theta_i t)^2}\left(1 - \frac{1}{(1+\theta_j t)}\right) + \frac{\theta_j}{(1+\theta_j t)^2}\left(1 - \frac{1}{(1+\theta_i t)}\right).$$

Table 2: Second-order inclusion probabilities for successive sampling design

i	j				
	1	2	3	4	5
1		0.0036335	0.0059265	0.0121894	0.0662504
2			0.0127577	0.0262163	0.1416418
3				0.0426846	0.2289937
4					0.4597060

Example 5. We compute second-order inclusion probabilities for Pareto πps sampling design for the same target inclusion probabilities vector given in the example 4, with population size $N = 5$ and sample size $n = 2$ elements. The control sum is $\sum_{i=1}^{N-1} \sum_{j=1}^{N} \tilde{\pi}_{i,j} = 1$, see Table 3.

Table 3: Second-order inclusion probabilities for Pareto πps sampling design

i	j				
	1	2	3	4	5
1		0.0033026	0.0053607	0.0112222	0.0746742
2			0.0112852	0.0234384	0.1517143
3				0.0374723	0.2357103
4					0.4458199

4.3 Conditional Poisson sampling case

The second-order inclusion probability of units i, j to be included in the sample $s, i \neq j$, can be derived similarly as in the univariate case, using Lemma 1 (Aires, 1999, p. 459) and by consideration of the equations,

$$\tilde{\pi}_{i,j} = \frac{p_i p_j S_{n-2}^{N-2}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_{j-1}, p_{j+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}. \tag{8}$$

Second-order inclusion probabilities are calculated in the same way as first-order inclusion probabilities. The program was developed with statistical package R. The input for this program is a vector of unconditional Bernoulli probabilities $p = (p_1, p_2, ..., p_N)$. As a result program gives computed second-order inclusion probabilities $\tilde{\pi}_{i,j}$.

Example 6. For given vector $p = (0.1, 0.2, 0.3, 0.5, 0.9)$, second-order inclusion probabilities $\tilde{\pi}_{i,j}$ are calculated, where population size $N = 5$ and sample size $n = 2$ elements. The results are shown in Table 4. Notice that control sum is $\sum_{i<j} \tilde{\pi}_{i,j} = n(n-1)/2 = 1$.

Table 4: Second-order inclusion probabilities for conditional Poisson sampling design

i	j				
	1	2	3	4	5
1		0.0016264	0.0027881	0.0065056	0.0585502
2			0.0062732	0.0146375	0.1317379
3				0.0250929	0.2258364
4					0.5269517

5 Conclusions

Simulation results show that inclusion probabilities for all sampling designs are close, but they do not coincide. We can see that using order sampling design with fixed order distribution shape the exact inclusion probabilities were approximated quite good. The differences can be explained by approximate numerical integration used for calculation of the inclusion probabilities, also by actual differences of those probabilities. The second-order inclusion probabilities differ more than the first-order inclusion probabilities. An approximate integration methods used for calculation of the inclusion probabilities requires long computer execution time .

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.

Krapavickaitė, D. (2012a). *Implementation of quality improvement actions for the labour force survey.* Statistics Lithuania.

Krapavickaitė, D. (2012b). *Order sampling with fixed distribution shape, calculation of the inclusion probabilities.* Manuscript.

Rosén, B. (1996). On sampling with probability proportional to size. *R&D Report. Research-Methods-Development* **1**, 1 – 25.

Rosén, B. (1998). On inclusion probabilities for order sampling. *R&D Report. Research-Methods-Development* **2**, 1 – 23.