CHALMERS | GÖTEBORG UNIVERSITY

Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs

N I B I A A I R E S

Department of Mathematical Statistics
CHALMERS UNIVERSITY OF TECHNOLOGY
GÖTEBORG UNIVERSITY
Göteborg, Sweden 2000

Thesis for the Degree of Doctor of Philosophy

Techniques to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs

Nibia Aires

CHALMERS | GÖTEBORG UNIVERSITY

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg, Sweden
Göteborg, March 2000

Techniques to Calculate Exact Inclusion
Probabilities for Conditional Poisson Sampling and
Pareto πps Sampling Designs
Nibia Aires
ISBN 91-7197-888-7

Doktorsavhandlingar vid Chalmers tekniska hgskola
Ny serie nr 1579
ISSN 0346-718X

Department of Mathematical Statistics
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Abstract

This thesis consists of five papers and treats two finite population sampling methods, viz. the Conditional Poisson and the Pareto πps sampling schemes. Both methods belong to a class of sampling schemes with unequal inclusion probabilities, commonly used in approximate probability to size sampling schemes.

Paper A addresses the problem of determining first and second order inclusion probabilities for both methods, which is a vital element in deriving linear estimators. Tools which consist of dynamic programming algorithms, are created to calculate these exact inclusion probabilities. They make it possible to compute the inclusion probabilities in reasonable execution time for small and moderate samples. Algorithms to adjust the parameters so that arbitrary desirable exact inclusion probabilities are achieved are also given. The results in this paper opened the possibility to use the exact Horvitz-Thompson and Yates-Grundy-Sen variance estimators for quite large samples for Conditional Poisson and for moderate samples in the Pareto case.

In Paper B, using those algorithms, a computer system was developed to compare and contrast the Conditional Poisson and Pareto πps sampling designs in terms of estimators of population totals, biases and variances. The computer programme produced an empirical comparison of both methods to check the convergence to the asymptotical inclusions. It also enabled adjustment of the parameters to obtain exact variances and make comparisons in these terms. The results from these studies show that the Pareto scheme approaches asymptotical inclusions faster than the Conditional Poisson, and that both methods are very similar in terms of second order inclusions for the adjusted procedures for moderate samples.

The second order inclusion algorithms for the Conditional Poisson Sampling design are generalised in Paper C to a recursive fast procedure to derive higher order inclusion probabilities of arbitrary order.

Paper D proves the existence and partial uniqueness of a set of scale parameters when exact inclusion probabilities are required for any order sampling of fixed distribution shape, a class of schemes of which Pareto πps sampling is a special case.

Lastly, Paper E reports on a thorough study of approximation accuracy for Pareto inclusion probabilities, aiming at practical use recommendations, for using asymptotically motivated approximations. Numerical results in this study are presented in Appendix 1 and 2.

Keywords: Sampling Theory, Conditional Poisson Sampling, Pareto *πps* Sampling, Order Sampling, Numerical Integration, Algorithms

AMS 1991 subject classification: 62D05, 65U05, 62-04

Acknowledgements

This thesis is the outcome of five years of post graduate education which involved taking classes, consulting, teaching and researching. Involvement in numerous activities over an extended period of time results in many people to acknowledge. My friends and colleagues Professor Elisabeth Svensson, Professor Jacques de Maré, Verónica Gaspes and Tommy Norberg have supported me and encouraged me during this journey. Professor Holger Rootzén was very helpful with the layout and putting things together. I received invaluable assistance on the algorithms and programming from Björn von Sydow and Thomas Ericsson, as well as from Andreas Jonasson for his consistently timely responses to problems. I had many interesting discussions with Professor Bengt Rosén preparing Paper E as well as with Johan Jonasson preparing Paper D. I am grateful to the people at the administration of the Department who have always been very helpful and supportive.

I have been pleased to be part of a university and department which is committed to making changes to respect to how mathematics is interacting with society and industry. Their efforts with the Centre for Applied Mathematics and Statistics and the Stochastic Centre are examples of this type of integration.

I also want to acknowledge the efforts of the school to bring more women into technology and mathematics. They have begun this journey and are making progress but it is a big challenge and I encourage them to continue pushing forward. I have realized to make progress with these major changes intelligence is necessary but not sufficient, you also need wisdom. I have been fortunate to work closely with three professors who have both. Professor Peter Jagers and Professor Sture Holm have given me valuable guidance. My supervisor Professor Olle Nerman is both and excellent researcher and has been very flexible which has allowed me to combine my research with my family life.

Even with this flexibility there have of course been some strains on my family life and I want to thank my husband Delvo and my daughters Adriana, Andrea, Natahi and Melina for the support and energy they have always given me. I want also to express my gratitude to my parents, my brothers and friends for being so close to me no matter the geographical distance.

Nibia Aires
Göteborg, March 2000

Contents

List of Papers

This thesis is composed of the following papers:

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

- **Paper B.** N. Aires, Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs, *Journal of Statistical Planning and Inference*, **82**, No. 2, pp. 1-15, 2000.

- **Paper C.** N. Aires, Inclusion Probabilities of Higher Order for Conditional Poisson Sampling, Chalmers University of Technology and Göteborg University, 1999.

- **Paper D.** N. Aires, J. Jonasson, O. Nerman, Order Sampling Design with Prescribed Inclusions, submitted to *Scandinavian Journal of Statistics*, 1999.

- **Paper E.** N. Aires, B. Rosén, On Inclusion Probabilities for Pareto πps Sampling, *Chalmers University of Technology and Statistics Sweden*, 2000.

In addition, the numerical results obtained in Paper E are available in Appendices 1 and 2.

Summary of the Papers

This thesis treats the problem of two sampling methods, viz. the Conditional Poisson (CPS) and the Pareto πps sampling (PPS) schemes which belong to a class of sampling methods with unequal inclusion probabilities. The study addresses the problem of determining first and second order inclusion probabilities, which is a vital element in deriving linear estimators.

Introduction

Information is a key issue in society. Knowledge about opinions in politics, consuming habits, preferences in sports and the arts, etc., are crucial for decision making and development in those areas. The theory of survey sampling offers tools and effective methods to obtain such information with a reliability that can be expressed and is based on data collection. In the frame of mathematical laws, a partial investigation of the finite population in question is enough for statistical inference about the population as whole. A firm basis for this kind of procedure may be achieved by probability sampling, which introduces an element of randomness into the sampling procedure. There are many probability sampling methods. But a good sampling scheme is one that is simply implemented, that leads to good estimation precision and that provides good variance estimation properties.

Brewer *et. al.* (1983) define two parts in the sampling strategy: the selection procedure, which deals with the way of choosing a sample from the population, and the estimation procedure. The latter states how inference will be carried out from the sample to the population. Furthermore, the estimation procedure may be also classified as enumerative or analytical, depending on the aim of the study. The purpose of enumerative inference is to describe the population. The object may be for instance to calculate population means, totals, proportions and ratios. On the other hand the purpose of an analytical inference is to explain the parameters by the use of a model with its own probability structure. For enumerative inference, the only relevant probability structure is the one determined by the manner in which the sample is selected.

As new methods for analysing data develop, it remains a fundamental requirement of good survey practice that a measure of precision, commonly the variance of the estimator, be provided for each estimate. In the derivation of an estimator of variance both estimator and sampling design must be taken into account.

As we shall see in this thesis, it is often rather difficult to find exact first and second order inclusion probabilities in order to construct unbiased esti-

mators and estimators of their variances. This is true for many interesting sampling designs, e.g. for CPS and PPS.

The historical perspective

The theory for independent random sampling was developed by Bernoulli two centuries ago; the theory of stratification, due to Poisson, also dates back many years. Notwithstanding the theory and application of survey sampling took a while to be established, it has developed dramatically in the last several decades.

In 1895, the subject of the survey sampling method was placed for the first time on the agenda of a session of the International Statistical Institute, cf. Dalenius (1957). By that time, sample surveys were distrusted by statisticians and non-statisticians alike, therefore they were used only occasionally. Survey procedures were classified into non-random and random sampling. Only in some rare cases matters of representativity were discussed. Most applications based on random sampling were systematic sampling from official records. There was a strong resistance against the use of sample surveys in traditional fields. The development in survey sampling is due to the development in industry, for instance, the forest industry in Sweden, and the social problems that arose in relation with these changes. In addition, economical reasons were an impediment to use total surveys in social studies. Also, the growth of the monetary economy created an interest in analysing consumer habits by the use of survey sampling.

It was only around 1936, when for first time, a partial population census in Sweden was planned to achieve maximum reliability, cf. Dalenius (1957). New methods such as stratified sampling, and the concept of optimally designed sampling, Neyman (1938), appear in practice in 1950, when there were examples of surveys planned so as to reach balance between cost and precision.

The use of unequal probabilities in sampling was first suggested by Hansen and Hurwitz (1943), who showed that the use of unequal probabilities generated more efficient estimators of the population total than those of equal probabilities. Madow (1949) proposed the use of systematic sampling with unequal probabilities to avoid the possibility of units being selected more than once. After this a large number of alternative sampling procedures were suggested, cf. Cochran (1977).

Horvitz and Thompson (1952) developed a general theory of sampling with unequal probabilities without replacement, based on the Horvitz and Thompson estimator used to estimate the population total. It is well known that the Horvitz and Thompson estimator is unbiased. Roy and Chakravarti (1960) showed that, for a given sampling scheme, there is no member of

the class of all homogeneous linear unbiased estimators of population total which has smaller variance than the Horvitz-Thompson estimator.

Notation

In statistical literature there are different definitions pertaining to sampling, but we think of a sample s as a subset of the population U, which is a given finite set. Some sampling designs may allow multiplicity of the units, which is called sampling with replacement.

Moreover, suppose that we are interested in an estimation of the population total $t_y = \sum_U y_i$ of a real value variable y, often called the study variable, or the population mean of the study variable y, $\overline{y} = \sum_U y_i/N$, where N is the number of elements in U. We assume that the values $y_i,\ i \in U$, are only known for the elements in the sample s selected from the population U. In this case t_y and $\overline{y}$ are examples of finite population parameters and the subset s is used to calculate estimates of t_y and $\overline{y}$.

The sample size, denoted by n is the number of elements in s, and the probability of selecting a sample s under the selection scheme in use will be denoted by $P(s)$. The latter are nonnegative numbers satisfying

$$\sum_s P(s) = 1,\ s \subset U.$$

The probability distribution $\{P(s),\ s \subset U\}$ is also called the sampling design and it determines the statistical properties of the random quantities calculated from the sample.

At the same time the sampling procedure in use assigns to each element i in the population U a probability to be included in the sample s, usually called the probability of inclusion. These inclusions are very important in the calculation of linear estimators. We define $\pi_i = P(i \in s)$ and $\pi_{ij} = P(i \in s, j \in s)$, thus we shall call π_i and π_{ij} the first- and second-order inclusion probabilities respectively.

The Horvitz-Thompson estimator of the population total is defined as,

$$\hat{t}_y = \sum_{i \in s} y_i/\pi_i.$$

This estimator is, provided all $\pi_i > 0$, the only unbiased estimator in the class of estimators for which the same weight is attached to a particular population unit whenever it is selected, cf. Horvitz and Thompson (1952).

Sampling with unequal probabilities

Naively and traditionally all elements were sampled with equal inclusion probabilities. However, one of the breakthroughs in sampling theory was

the realization that more accurate estimates can be obtained by assigning different inclusion probabilities to the elements in the population, that is to sample with unequal inclusion probabilities. If each element has the possibility to be chosen only once, the sample is said to be selected without replacement and the procedure is denoted by πps. A general theory for selection with this type of probabilities is presented in Horvitz and Thompson (1952). In this paper the Horvitz-Thompson estimator of the population total described above, is introduced. The variance of this unbiased estimator is,

$$V(\hat{t}_y) = \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

An unbiased estimator of $V(\hat{t}_y)$ is given by

$$\hat{V}(\hat{t}_y) = \sum_{i,j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

given that $\pi_{ij} > 0, \forall\, i \neq j \in U$. Provided that the sample size is fixed, an alternative expression for the variance of the estimator $\hat{t}_y$, derived independently by Sen (1953) and Yates and Grundy (1953), is

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$

This formula is valid only if the number of units in the sample is fixed. An alternative natural unbiased estimator of $V(\hat{t}_y)$ is then

$$\hat{V}(\hat{t}_y) = -\frac{1}{2} \sum_{i,j \in s} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

given that all $\pi_{ij} > 0, \forall\, i \neq j \in U$.

Moreover, it follows from the Sen-Yates-Grundy formula for $V(\hat{t}_y)$ that if all y_i are exactly proportional to the corresponding π_i and the number of units in the sample is fixed, the variance of the Horvitz-Thompson estimator is zero. An important application of this idea occurs when for each individual i a positive quantity x_i is known and it is believed that x_i is approximately proportional to a study variable y_i. In order to use this information to get good estimators of the total, $t_y = \sum_U y_i$, the choice of the sample s can be made using inclusion probabilities π_i approximately proportional to x_i. This kind of procedure is called probability proportional-to-size sampling.

In some cases, it may not be possible to select a sample based on strictly proportional inclusion probabilities π_i, simply by taking, $\pi_i = nx_i/\sum_U x_j$, $i = 1, ..., N$ since one or more of the $\pi's$ may exceed 1. This obstacle is usually circumvented in practice by introducing a "take all" stratum of units with the largest sizes. We shall assume in the sequel that $nx_i < \sum_U x_j,\ i = 1, ..., N$.

Conditional Poisson Sampling (CPS)

Poisson Sampling is an unequal probability sampling design with random size. To carry out a Poisson sampling, first assign a probability of inclusion to each population unit. Then perform N Bernoulli trials using these probabilities to determine whether or not the corresponding unit is to be included in the sample. All units for which the trials have been successful constitute the sample.

CPS is a sampling method with varying inclusion probabilities and fixed sample size. It may be defined as Poisson sampling conditioned by the requirement that the sample s belong to a subclass of samples.

Hájek (1964) makes an extensive study of Rejective sampling, which may be regarded as Poisson sampling conditioned that the sample size equals n and thus may be called CPS. He presents basic facts about the method and derives conditions for asymptotic normality of the estimators. He also introduces approximation formulas for calculation of first and second order inclusion probabilities, as well as an approximation formula for adjusting the parameters $p_i = nx_i/\sum_U x_j,\ i = 1, ..., N$, when the exact inclusions π_i are given. In a triangular setting, with a sequence of sample procedures from a sequence of populations, he proves that CPS yields probabilities of inclusion such that

$$\pi_i/p_i \to 1,\ i = 1, ..., N$$

as $N \to \infty$, uniformly in i, provided that $\sum_{i=1}^{N} p_i(1 - p_i) \to \infty$. CPS also maximises the entropy, which is a measure of spread of the sampling design $P(s)$, given by

$$-\sum_{s \subset U} P(s) \ln P(s),$$

in the class of sampling schemes with the same first order inclusions, see Hájek (1981), page 29. The distribution with maximal entropy in a class of distributions is the "most random" in that class, which is a desirable condition for a sampling scheme. In other words, this measure informs about how the population is represented in the sample.

A CPS may be realised by N independent Bernoulli trials determining whether the individual under consideration is to be included in the sample

s or not. In fact, the N trials form one experiment. Any experiment that results in other than n out of the N individuals being picked is rejected. One performs independent experiments sequentially until one of the experiments results in n out of the N individuals being picked.

Order Sampling

Order sampling schemes form a general class of sampling procedures with varying probabilities. Each unit within the population is assigned an independent but not equally distributed random variable, called a *ranking variable*. To choose a sample of size n without replacement, first these ranking variables are realized and then the units that have the n smallest ranking values constitute the sample. Rosén (1997) introduces the method and the notion of order Sampling with Fixed Distribution Shape. He states that by varying this distribution shape, a wide class of varying probabilities sampling schemes is obtained.

Ohlsson (1990, 1995) introduces sequential Poisson sampling, which is order sampling with uniform ordering distributions. The method which is a πps scheme has been used in the Swedish Consumer Price Index survey system since 1990.

Asymptotic results have been obtained for successive sampling by Hájek (1981) and Rosén (1972). Successive sampling is equivalent to order sampling with exponential ordering distributions.

Saavedra (1995) and Rosén (1997b), independently, introduced a particular order sampling scheme, viz. the Pareto πps procedure. The ranking variables Q_i, in this scheme, have the standard Pareto distribution function $F_i(t) = \theta_i t/(1+\theta_i t)$ with parameters $\theta_i > 0$, for $i = 1, ..., N$. Parallel to this θ parametrisation we use an alternative set of parameters which are more directly coupled to the inclusion probabilities: $\lambda_i = F_i(1) = \theta_i/(1+\theta_i)$, $i = 1, ..., N$. This is motivated by the fact that λ_i approximates the inclusion probabilities in case $\sum_U \lambda_i = n$, cf. Rosén (1997b).

Rosén (1997b) also studies the asymptotic distributions of linear statistics for order sampling with fixed distribution shape. He also proves that the Pareto scheme is optimal among Order sampling schemes with Fixed Distribution Shape, in the sense that it minimises the estimator variances asymptotically.

Outline of the papers

The present thesis contains five papers which treat the Conditional Poisson Sampling (CPS) and Pareto πps Sampling (PPS) designs. Both methods belong to the class of probability proportional-to-size sampling schemes.

They also have good sampling properties and they present interesting similarities. However, exact expressions for first order inclusion probabilities are intricate and matters become still worse for second order quantities. Previous studies on Horvitz-Thompson estimator biases and variances have been done for both methods based on asymptotical approximations due to this problem.

Paper A. Algorithms to Find Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs, N. Aires

In this paper we create the tools to calculate the exact inclusion probabilities for both methods. The tools consist of algorithms that make it possible to compute first and second order inclusion probabilities in reasonable execution time for small and moderate samples, using computers.

To give an idea of the numerical calculations performed in the paper we first introduce the inclusion probabilities in terms of the Poisson parameters. The problem consists in calculating the quantities,

$$\breve{\pi}_i = \breve{\pi}_i(p) = P(i \in s \mid \mid s \mid = n) = \frac{\sum_{s \in A_n^i} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)}{\sum_{s \in A_n} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)},$$

$i = 1, ..., N$, where $\mid s \mid$ is the cardinality of s, A_n denotes the class of all possible samples of size n and A_n^i the set of elements of A_n containing i; for first order inclusion probabilities. And analogously for second order inclusions,

$$\breve{\pi}_{ij} = P(i,j \in s \ \mid \mid s \mid = n) = \frac{\sum_{s \in A_n^{ij}} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)}{\sum_{s \in A_n} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)},$$

where A_n^{ij} is the set of elements of A_n containing both i and j, $i, j = 1, ..., N$. Adopting a dynamic programming algorithm makes it possible to calculate $\breve{\pi}_i$ and $\breve{\pi}_{ij}$ quite fast. By separating terms according to whether the element N is or is not in s and by defining

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

with $N = 0, 1, 2, ...$ and $n = 0, ..., N$; S_n^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 $S_0^N = (1-p_1)(1-p_2)\cdots(1-p_N)$ and $S_N^N = p_1 p_2 \cdots p_N$.

Then, the inclusion probability of any unit i, $i = 1, ..., N$, can be written as,

$$\breve{\pi}_i = \frac{p_i S_{n-1}^{N-1}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}.$$

Second order inclusion probabilities for CPS can be calculated using the same principle. However, a more effective algorithm to calculate bivariate inclusions is presented in the paper. Let $\breve{\pi}_{i\setminus j}$ denote the probability that element i but not element j belongs to the sample s and define $\gamma_i = p_i/(1-p_i)$. Then based on the relations

$$\begin{cases} \breve{\pi}_i = \breve{\pi}_{i\setminus j} + \breve{\pi}_{ij} \\ \\ \breve{\pi}_j = \breve{\pi}_{j\setminus i} + \breve{\pi}_{ij}, \end{cases}$$

it is easy to see that

$$\breve{\pi}_{ij} = \frac{\gamma_i \breve{\pi}_j - \gamma_j \breve{\pi}_i}{\gamma_i - \gamma_j}$$

for the case $\gamma_i \neq \gamma_j$. For the case $\gamma_i = \gamma_j$, second order inclusions probabilities are calculated using

$$\breve{\pi}_{ij} = \frac{(n-1)\breve{\pi}_i - \sum_{j:\gamma_j \neq \gamma_i} \breve{\pi}_{ij}}{k_i}.$$

where k_i is the number of elements $j \neq i$ such that $\gamma_j = \gamma_i$ or equivalently $p_j = p_i$. Thus, once we have the first order inclusion probabilities we need virtually no time to get the second order inclusions as well.

For the PPS procedure, first the λ-parameters are calculated by taking,

$$\lambda_i = n x_i / \sum_U x_j, \ i = 1, ..., N.$$

We assume as usual that all $\lambda_i < 1$.

Then we consider independent ranking variables $Q_1, Q_2, ..., Q_N$ with corresponding Pareto distribution functions $F_1, F_2, ..., F_N$. The units with the n smallest Q-values constitute the sample s.

Furthermore, the probability of element N belonging to the sample s is

$$\begin{aligned} \tilde{\pi}_N &= P(N \in s) = P(Q_{(n)}^{N-1} > Q_N) \\ &= \int_0^\infty (1 - F_n^{N-1}(t)) f_N(t) dt, \end{aligned}$$

where $Q_{(n)}^{N-1}$ is the n:th order statistic among $Q_1, Q_2, ..., Q_{N-1}$ with distribution function F_n^{N-1} and $f_N(t)$, $i = 1, ..., N$ is the density of Q_N. The integrand in this inclusion probability integral is calculated by a dynamic programming algorithm, which is in turn combined with numerical integration procedures to give the inclusion. In fact $F_n^N(t)$, $N = 1, 2, ...$, $n = 1, ..., N$, satisfy the recursive equation:

$$F_n^N(t) = F_n^{N-1}(t) + F_N(t)[F_{n-1}^{N-1}(t) - F_n^{N-1}(t)].$$

Thus $F_n^N(t)$ may be calculated recursively using the observation that $F_0^N(t) = 1$, for all N *and* t. A similar formula can be derived for any other π_i by rearranging the order of the Q_i variables in the above formula.

In the same way, second order inclusion probabilities may be calculated by

$$\begin{aligned} \tilde{\pi}_{N-1,N} &= P(N-1 \in s, N \in s) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N)) \\ &= \int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1},Q_N)}(t)dt, \end{aligned}$$

using the same algorithm to calculated $F_{n-1}^{N-2}(t)$ and combining with numerical integration procedures.

First order inclusion probabilities can be used in the Horvitz-Thompson estimator formula to obtain unbiased estimators. Second order inclusion probabilities can be used to calculate the exact variance of Horvitz-Thompson estimators in simulation studies where the response variable is known, or for variance estimation.

In this paper we also present algorithms to adjust the parameters. The adjusting procedures provide the parameters to achieve desirable exact inclusion probabilities. For CPS, we assume that the exact inclusion probabilities z_i are given and the task is then to obtain the Poisson parameters p_i. The problem can be expressed as a non-linear equation system with N unknown variables and $N + 1$ equations of the form,

$$\begin{cases} \breve{\pi}_i(p_1, ..., p_N) - z_i = 0, \quad i = 1, ..., N \\ \\ \sum_{i=1}^N p_i - n = 0. \end{cases}$$

We solve this equation system numerically by using a built-in function in Matlab based on the Gauss-Newton method and denote the solution p_i^A. This has earlier been shown to exist and be unique in Dupačová (1979). For PPS the non-linear equation system to solve the adjusting problem is

$$\begin{cases} \tilde{\pi}_i(\lambda_1, ..., \lambda_N) - z_i = 0, \quad i = 1, ..., N \\ \\ \sum_{i=1}^N \lambda_i - n = 0. \end{cases}$$

Using the same algorithm as in the CPS case to solve this equation system requires long execution time for the PPS scheme. Therefore we develop a more simple idea to get an alternative adjusting procedure that converges faster to the adjusted values λ_i^A. The procedure consists in iterating

$$\lambda^A(k+1) = \lambda^A(k) + (z - \tilde{\pi}(\lambda^A(k)))$$

until

$$\max \left| \frac{\tilde{\pi}_i(\lambda^A(k))}{z_i} - 1 \right| \leq 10^{-4},$$

given the start value $\lambda^A(0) = z$. After each iteration we adjust the $\tilde{\pi}_i(\lambda^A(k))$ values slightly by normalising,

$$\tilde{\pi}_i(\lambda^A(k)) = \left(\frac{\tilde{\pi}_i(\lambda^A(k))}{\sum_j \tilde{\pi}_j(\lambda^A(k))} \right) \cdot n.$$

This algorithm was implemented applied in some of our larger sample size examples obtaining good results in terms of precision and execution time.

These calculations facilitate to check the asymptotic behaviour of the inclusion probabilities and the possibility to use the latter in survey sampling. Some results can be found in Paper B where an analogous algorithm for the Poisson case also was implemented and for the PPS case in Paper E.

Paper B. Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs, N. Aires

In this paper, we compare both methods in terms of calculations of relative biases and variances of the estimators of the population totals, by using the algorithms in Paper A. In the first part of this paper, we present calculations of first and second order exact inclusion probabilities for both sampling schemes for small, moderate and large population sizes. A computer system was developed to make comparisons between CPS and PPS designs in terms of estimators of population totals, biases and variances. The algorithms in Paper A were re-implemented in Fortran and the use of the routines for numerical integration of this programming language made substantial improvements in the accuracy of the results and the execution time of the programs. The program calculates first and second order inclusion probabilities as well as the adjusted inclusion probabilities for both schemes. The second order inclusion probabilities for the adjusted schemes are also calculated. Since a file which provides a study variable and auxiliary information is coupled to these routines, the real variance of $\hat{t}_y$ is computed

using the Yates-Grundy-Sen formula. The asymptotic variance, given by

$$AV^A = \frac{N}{N-1}\sum_{i=1}^{N}\left(\frac{y_i}{z_i} - \frac{\sum_j y_j \frac{p_j^A}{z_j}(1-p_j^A)}{\sum_j p_j^A(1-p_j^A)}\right)^2 \cdot p_i^A(1-p_i^A)$$

is also calculated for the CPS scheme. This formula coincides with Hájek's asymptotic variance formula, except for the factor $N/N-1$ which is negligible when N is large, cf. Hájek (1964). The program also computes the asymptotic variance for the adjusted PPS scheme, cf. Rosén (1997b), given by

$$AV^A = \frac{N}{N-1}\sum_{i=1}^{N}\left(\frac{y_i}{z_i} - \frac{\sum_j y_j \frac{\lambda_j^A}{z_j}(1-\lambda_j^A)}{\sum_j \lambda_j^A(1-\lambda_j^A)}\right)^2 \cdot \lambda_i^A(1-\lambda_i^A).$$

The results show that for the PSS design the inclusion probabilities converge faster to the given inclusion probabilities than for the CPS scheme.

Moreover, to compare both methods we construct a measure of variance dissimilarities, a bound for this measure and a bound for the relative bias. The program calculates the relative absolute bias (only in case of positive signs variables) and its bound for PPS by

$$\tilde{B} = \left|\frac{\tilde{E}(\hat{t}_y) - t_y}{t_y}\right| = \left|\frac{1}{t_y}\sum y_i\left(\frac{\tilde{\pi}_i}{\lambda_i} - 1\right)\right| \leq max\left(\left|\frac{\tilde{\pi}_i^R}{z_i} - 1\right|\right) = \tilde{B}_U.$$

In the same way, the relative bias B and its bound are calculated for the CPS case, $B \leq max\left|(\pi_i^R/z_i) - 1\right| = B_U$. The program also computes the variance dissimilarity ψ and its bound ψ_U for the two unbiased sampling schemes, by using the formula,

$$\psi = \frac{|\,\tilde{V}^A(\hat{t}_y) - V^A(\hat{t}_y)\,|}{V^A(\hat{t}_y)} \leq max_{i\neq j}\frac{|\,\tilde{\pi}_{ij}^A - \pi_{ij}^A\,|}{|\,\pi_{ij}^A - z_i z_j\,|} = \psi_U.$$

$\tilde{V}^A$ and $\tilde{\pi}_{ij}^A$; V^A and π_{ij}^A are the real variance and the second order inclusion probabilities for PPS and CPS respectively using the adjusted values. Calculations of these were performed for different population sizes. The results obtained corroborate the similarity of both schemes. In all given examples it is observed that even if the asymptotic variances are very close in both schemes, the asymptotic variance for PPS design approximates the real variance better and this real variance is slightly smaller for CPS.

Paper C. Inclusion Probabilities of Higher Order for Conditional Poisson Sampling, N. Aires

We generalise two methods to calculate inclusion probabilities of higher order for CPS design based on the idea presented in Paper A to derive bivariate inclusion probabilities. The first method, which is more time consuming, is based on the same principle to calculate first order inclusion probabilities for CPS. On the other hand, a more effective algorithm computes recursively higher order inclusion probabilities using the Poisson parameters and the exact inclusion probabilities. For instance to calculate 3rd order inclusion probabilities the algorithms takes into account the Poisson parameters p_i and the second order inclusions $\breve{\pi}_{ij}$. Thus the probability that elements i, j and k are included in the sample s denoted by $\breve{\pi}_{ijk}$ may be calculated by

$$\breve{\pi}_{ijk} = \frac{\gamma_j \breve{\pi}_{ik} - \gamma_k \breve{\pi}_{ij}}{\gamma_j - \gamma_k}$$

where $\gamma_j = p_j/(1-p_j)$, in the case $\gamma_j \neq \gamma_k$. Special formulas are presented in the paper for the case $\gamma_j = \gamma_k$, for the third order inclusions as well as for the general case.

These inclusion probabilities are of interest for calculation and estimation of moments of higher order, e.g. to facilitate corrections on the coverage accuracy of the confidence interval of the estimator.

Paper D. Order Sampling Design with Prescribed Inclusions, N. Aires, J. Jonasson, O. Nerman

In Paper D we prove the conjecture that the parameters in Order Sampling can be arbitrarily prescribed for any order sampling with fixed distribution shape. As we mentioned before, methods were provided to compute exact first and second order inclusion probabilities numerically when the distribution shape is of the Pareto type. Procedures were also provided for this case to adjust the parameters to get predetermined inclusion probabilities. However, the existence of the latter was not really proved in an arbitrary case, but it was numerically derived in many examples.

In this paper we prove the existence and partial uniqueness of a solution for the latter problem, in general for any order sampling of fixed distribution shape. As a special result follows that the equation system used to find the adjusted Pareto values λ^A in Paper A has exactly one solution. The argument follows the same line as for a corresponding earlier proof of a similar theorem for CPS procedures.

Paper E. On Inclusion Probabilities for Pareto πps Sampling, N. Aires, B. Rosén

Previous studies on Pareto schemes are based on asymptotic considerations. As a consequence, the factual inclusion probabilities deviate to some extent from the prescribed ones. Methods to calculate exact inclusion probabilities for this procedure introduced in Paper A, make it possible not only to obtain unbiased estimators and exact variances, but also to check the asymptotic behaviour of the inclusion probabilities. Moreover, there is a close connection between approximation accuracy for inclusion probabilities and estimator bias. The magnitude of the latter is a question of practical relevance.

Paper E reports on a thorough study of approximation accuracy for Pareto inclusion probabilities, aiming at practical use recommendations of the asymptotic theory. The paper presents conditions which ensure that the estimator bias is practically negligible. The chief tool is the computation algorithm for Pareto πps inclusion probabilities in Paper A. Some basics on Pareto sampling are introduced as well as various measures of approximation accuracy. The central measure of approximation goodness is the maximal absolute relative bias (in Paper B called $\tilde{B}$) for inclusion probabilities given by

$$\Psi = \max_i \left| \frac{\pi_i}{\lambda_i} - 1 \right|, i = 1, 2, ..., N.$$

The name is motivated by the fact that for a non-negative study variable $\mathbf{y}$, the absolute relative bias for the population total $\tau(\mathbf{y}) = y_1 + ... + y_N$ is

$$\left| \frac{E[\hat{\tau}(\mathbf{y})] - \tau(\mathbf{y})}{\tau(\mathbf{y})} \right| \leq \Psi.$$

This bound is often conservative (but can always be attained) and depends on the population size, the sample size and the size values. The size values conform a vector of auxiliary information, denoted by $s = (s_1, ..., s_N)$, $s_i > 0$, $i \in U$, which typically is correlated to the study variable and relates to the inclusion probabilities as follows

$$\lambda_i = n \cdot s_i / \sum_{j=1}^{N} s_j, \ i = 1, 2, ..., N.$$

These size measures are also normed so that the average size is 1,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} s_i = 1.$$

Moreover, the study is based on certain size value patterns which play an important role in the numerical results. These patterns, named "boundary", "middle" and "even", take into account the range and the shape of the size values. We denote by γ and δ the smallest and largest size value respectively for a given pattern. The paper addresses three main questions, viz.

- How large is Ψ for a specific sample size n?
- For which samples sizes n, is Ψ less than a specified value β?
- Will the estimator bias be negligible in a particular sampling estimation situation?

Answers to these questions are presented in the paper. Conclusions concerning safe practical use of Pareto sampling are also exhibited in the paper. For instance, it is shown that the accuracy improves as population size increases. From the results we also conclude that for almost all situations which run up in practice, the Pareto πps can safely be used.

Conclusions

The algorithms introduced in this thesis make it possible to calculate exact inclusion probabilities for CPS and PPS. These results in turn enables the use of both sampling schemes in combination with unbiased estimators in small and moderate samples (up to say two hundred, chosen among a population of thousands of elements seems feasible with a more refine computer program). The results also offer the possibility to use the exact Horvitz-Thompson and Yates-Grundy-Sen variance estimators for equally large samples for CPS and for moderate samples in the Pareto case.

The implementation also creates the possibility to compare, in an empirical way, both methods to check the convergence to the asymptotical inclusions. Moreover these new methods enable us to adjust the parameters to obtain exact variances and make comparisons in these terms. The results from these studies show that both sampling schemes are very similar in terms of second order inclusion probabilities for the adjusted procedures for moderate samples.

The algorithms and the implemented programs were also used to check asymptotical results in approximation studies for moderate sample sizes. The results in those indicated that the approximations are very good for larger samples.

Topics of further work

One interesting topic that remains to be studied, is related with Hájek's approximation formulas for first and second order inclusion probabilities

of CPS. Thus Hájek namely suggested refinements of those approximation formulas, cf. Hájek (1964), it would be of great interest to compare the exact inclusions with these refined approximations in a future, systematic study of the effects of population sizes and inclusion patterns.

Next, we wish to document and rewrite the programmes into a system aimed at general use. In this work we also wish to include the implementation of the algorithm for calculating inclusion probabilities of higher order for the CPS scheme.

Furthermore, we desire to implement a more general order sampling programme for an arbitrary distribution shape for calculating first and second order inclusion probabilities. We also wish to find a fast adjusting algorithm, analogous to the Pareto procedure, for cases with distribution support on $[0, \infty)$ and for the uniform distribution case, i.e. Sequential Poisson Sampling. The implementation of these tools could make the method more useful in practice and maybe replace the approximation methods.

Finally, a formal proof of the convergence and rate of convergence of the heuristic algorithm used for the adjusted procedure is desirable in a future study.

References

Brewer K.R.W., Hanif, M. (1983). *Sampling with Unequal Probabilities*, Springer-Verlag, New York.

Dalenius, T. (1957). *Sampling in Sweden*, Contributions to the methods and theories of sample survey practice. Almqvist and Wicksell, Stockholm.

Cochran, W. (1977). *Sampling Techniques*, Third Edition. John Wiley & Sons, Singapore.

Dupačová, J. (1979). A note on Rejective Sampling, Contributions to Statistics, Jaroslav Hájek Memorial Volume, Reidel, Holland and Academia, Prague, 71-78.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population. *Annals of Mathematical Statistics*, **35**, 1491-1523.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Hansen, M.H., Hurwitz, W.N. (1943). On the theory of sampling from a finite population. *Annals of Mathematical Statistics*, **14**, 333-362.

Madow, W.G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, **20**, 333-354.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, **20**, 333-354.

Ohlsson, E. (1990). Sequential Poisson sampling from a business register and its application to the Swedish consumer price index. Statistics Sweden R & D Report 1990:6.

Ohlsson, E. (1995). Sequential Poisson Sampling. Report No. 182. Institute of Actuarial Mathematics and Mathematical Statistics. Stockholm University, Sweden.

Roy, J., Chakravarti I.M. (1960). Estimating the Mean of a Finite Population, *Annals of Mathematical Statistics.*, **31**, 392-398.

Rosén, B. (1972). Asymptotic Theory for Successive Sampling with Varying Probabilities without Replacement I and II, *Annals of Mathematical Statistics.*, **43**, 373-397 and 748-776.

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. (1998). On Inclusion Probabilities of Order Sampling. R&D Report, Research - Methods - Development No. 2, Statistics Sweden.

Saavedra, P. (1995). Fixed Sample Size PPS Approximations with a Permanent Random Number. *1995 Joint Statistical Meetings American statistical Association.* Orlando, Florida.

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

Paper A

Reprinted from METHODOLOGY AND COMPUTING IN APPLIED PROBABILITIES, Vol. 1:4, pp. 457-469, 1999, Algorithms to Calculate Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs, Aires N., ©2000, with kind permission from Kluwer Academic Publishers.

Algorithms to Find Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs

Nibia Aires

School of Mathematical and Computing Sciences,
Chalmers University of Technology,
S-412 96 Gothemburg, Sweden

Received December 1, 1998; Revised September 20, 1999, Accepted September 21, 1999

Abstract

Conditional Poisson Sampling Design as developed by Hajék may be defined as a Poisson sampling conditioned by the requirement that the sample has fixed size. In this paper, an algorithm is implemented to calculate the conditional inclusion probabilities given the inclusion probabilities under Poisson Sampling. A simple algorithm is also given for second order inclusion probabilities in Conditional Poisson Sampling. Furthermore a numerical method is introduced to compute the unconditional inclusion probabilities when the conditional inclusion probabilities are predetermined. Simultaneously, we study the Pareto πps Sampling Design. This method, introduced by Rosén, belongs to a class of sampling schemes called Order Sampling with Fixed Distribution Shape. Methods are provided to compute the first and second order inclusion probabilities numerically also in this case, as well as two procedures to adjust the parameters to get predetermined inclusion probabilities.

Keywords: Sampling Theory, Conditional Poisson Sampling, Pareto πps Sampling, Numerical Integration, Algorithms

AMS 1991 subject classification: 62D05, 65U05, 62-04

1. Introduction

Hájek (1964) studied the behaviour of the Horvitz-Thompson estimator under rejective sampling design of size n, which is equivalent to Poisson Sampling given the condition that the sample size equals n. We shall use the name Conditional Poisson Sampling (CPS) in the sequel. In the same paper, the author also showed that a central limit theorem holds for such a design and considered the relation between the inclusion probabilities under Poisson Sampling and the conditional inclusion probabilities. At the same time, he proposed approximation formulas for achieving the conditional inclusion probabilities as well as formulas for adjusting them. Later Dupačová (1979) showed that the inclusion probabilities in CPS may be arbitrarily prescribed. However, in both cases the resulting probabilities remain to be found in practice.

On the other hand, an asymptotic theory for the Pareto πps Sampling (PPS) design is given by Rosén (1997b). He shows that this scheme is asymptotically uniformly optimal among the schemes which have inclusion probabilities proportional to given size measures (πps) and belongs to a class of sampling schemes called order sampling with fixed distribution shape. But, as in the CPS case, the exact inclusion probabilities for the PPS scheme are to be found.

In the first part of this paper, algorithms to find numerical values of the exact inclusion probabilities for both sampling schemes are given along with a few examples. A method to compute the second order pair specific inclusion probabilities is also provided and compared for the CPS and the PPS schemes. Finally, we introduce numerical methods to adjust the parameters when the exact inclusion probabilities are given for both procedures.

The procedures described in this paper can be used in different ways. For instance, the exact first order inclusion probabilities can be used in the Horvitz-Thompson estimator formula to eliminate the estimator bias. Second order inclusion probabilities can be used to calculate the exact variance of Horvitz-Thompson estimators in simulation studies where the response variable is known or for variance estimation. On the other hand, the adjusting procedures provide the parameters to achieve desirable exact inclusion probabilities. The calculations may be achieved in reasonable execution time for small and moderate samples. In addition, the importance of these calculations is to make it possible to check the asymptotic behaviour of the inclusion probabilities and the possibility to use the latter in survey sampling; some results can be found in Aires (1998). Further research is in progress to systematically compare both πps procedures and to study the precision in earlier results where asymptotical approximations were used, cf. Aires (2000).

2. First order inclusion probabilities

2.1. Conditional Poisson Sampling (CPS) Design

Poisson Sampling is a method for choosing a sample s, of random size $|s|$, from a finite population U consisting of N individuals. Each individual i in the population has a predetermined probability p_i of being included in the sample s. A Poisson sample may be realised by using N independent Bernoulli trials to determine whether the individual under consideration is to be included in the sample s or not. Hájek (1981) showed that conditioning on the sample size n in a Poisson Sampling Design, yields the maximum entropy distribution of s among all sampling procedures of size n, with inclusion probabilities of the individuals equal to those of this CPS procedure. In fact, the N trials form one experiment. Any experiment that results in other than n out of the N individuals being picked is rejected. One performs sequentially independent experiments until one of the experiments results in n out of the N individuals being picked. Let A_n denote the class of all possible samples of size n and A_n^i the set of elements of A_n containing i, so that

$$\breve{\pi}_i \quad = \quad \breve{\pi}_i(p) = P(i \in s \mid |s| = n) = \frac{\sum_{s \in A_n^i} \prod_{j \in s} p_j \prod_{k \notin s} (1 - p_k)}{\sum_{s \in A_n} \prod_{j \in s} p_j \prod_{k \notin s} (1 - p_k)}, \tag{2.1}$$

$i = 1, ..., N$, are the inclusion probabilities of the individuals in the CPS procedure. It is hardly ever true that $\breve{\pi}_i = p_i$. Nevertheless a choice of the p_i's can be made by solving the following equation system, for any given set of probabilities z_i, $i = 1, ..., N$ satisfying $\sum_{i=1}^{N} z_i = n$,

$$z_i \quad = \quad \breve{\pi}_i(p), \; i = 1, ..., N. \tag{2.2}$$

Furthermore, the equation system (2.2) has a unique solution such that $\sum_{i=1}^{N} p_i = n$, as shown by Dupačová (1979). Hájek also showed that for large samples one can let p_i equal the desired conditional inclusion probabilities z_i to get a good approximation of the solution of equation (2.2). For moderate sample sizes he also suggested more elaborate approximative formulas in order to get more precise inclusion probabilities or vice versa approximative adjustment of the unconditional inclusions. However, it is of interest to achieve the exact inclusion probabilities to avoid biases of population estimators. How to solve (2.2) numerically will be discussed in Section 4.

An important application of the idea to sample with unequal inclusion probabilities occurs when for each individual i a positive quantity x_i is known and it is believed that x_i is approximately proportional to a study

variable y_i. In order to use this information to get good estimators of the total, $t_y = \sum_U y_i$, the choice of the sample s can be made using inclusion probabilities π_i proportional to x_i, and then the Horvitz-Thompson estimator $\hat{t}_y = \sum_{i \in s} y_i/\pi_i$ may be adopted to estimate the population total. This estimator is unbiased and has small variance, cf. Särndal *et. al.* (1992). In some cases, it may not be possible to select a sample based on strictly proportional inclusion probabilities π_i, simply by taking, $\pi_i = nx_i/\sum_U x_j$, $i = 1, ..., N$ since one or more of the $\pi's$ may exceed 1. This obstacle is usually circumvented in practice by introducing a "take all" stratum of units with the largest sizes. We shall assume from now on that $nx_i < \sum_U x_j$, $i = 1, ..., N$.

We return to the problem of calculating exact inclusion probabilities $\breve{\pi}_i$ for CPS, using the Bernoulli parameters p_i's. But let us first analyse the structure of equation (2.1).

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 $n \leq N$. Then the quantities*

$$S_n^N(p_1, ..., p_N) = \sum_{s \in A_n(N)} \prod_{i \in s} p_i \prod_{j \notin s} (1 - p_j) \tag{2.3}$$

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

$$S_0^N = (1 - p_1)(1 - p_2) \cdots (1 - p_N)$$

and

$$S_N^N = p_1 p_2 \cdots p_N.$$

Proof. The proof follows by separating terms according to whether N is or is not in s. ■

The inclusion probability of any unit i, $i = 1, ..., N$, can be written as,

$$\breve{\pi}_i = \frac{p_i S_{n-1}^{N-1}(p_1, ..., p_{i-1}, p_{i+1}, ..., p_N)}{S_n^N(p_1, ..., p_N)}, \tag{2.4}$$

so that Lemma 1 can be used to calculate it.

Example 1 The conditional inclusion probabilities are calculated by a computer program[1] using the recursion in Lemma 1. For any vector of unconditional Bernoulli probabilities $(p_1, ..., p_N)$ the program returns the vector of conditional inclusion probabilities $(\breve{\pi}_1, ..., \breve{\pi}_N)$. In Table 1 these inclusions are shown for the vector $p = (0.1\ 0.2\ 0.3\ 0.5\ 0.9)$, with $N = 5$ and $n = 2$. Notice that $\sum_{i=1}^{N} p_i = \sum_{i=1}^{N} \breve{\pi}_i = 2$.

Table 1: 1st. order inclusions, CPS.

p	$\breve{\pi}$
0.1	0.06947026022305
0.2	0.15427509293680
0.3	0.25999070631970
0.5	0.57318773234201
0.9	0.94307620817844
Sum: 2.0	2.00000000000000

Example 2 From a given vector of proportional inclusions, for a population of size $N = 284$ and sample size $n = 80$, the conditional inclusion probabilities were calculated. These resulting inclusions are very similar to the input vector, they differ in the fourth decimal. The execution time was ten minutes. An indication of the high quality of these numerically derived probabilities is that their sum is 80 using 8 decimals precision.

2.2. Order Sampling

Consider a population $U = \{1, ..., N\}$. To each unit i in the population is associated a probability distribution $F_i(t)$ with density $f_i(t)$, $0 \leq t < \infty$. To realize an Order Sampling scheme of sample size n, with $n < N$, we consider independent ranking variables $Q_1, Q_2, ..., Q_N$ with distribution functions $F_1, F_2, ..., F_N$. The units with the n smallest Q-values constitute the sample. The idea of Order Sampling originates from the special case where all Q_i are uniform in which case the inventor Esbjörn Ohlsson (1995), named the resulting procedure Sequential Poisson Sampling. The idea was further developed by Rosén (1997a), who also introduced the PPS scheme. Thus an important subclass is derived by the assumption that all F_i belong to the same scale family and in the case these distribution functions are defined as $F_i(t) = \theta_i t/(1+\theta_i t)$, $\theta_i > 0$, we have a PPS procedure. PPS based

[1]The programs used in the examples in this paper are written in Matlab (ver. 5.2) for the Unix system.

Horvitz-Thompson estimators are asymptotically uniformly optimal among order sampling schemes with inclusion probabilities proportional to given size measures and with fixed distribution shape, cf. Rosén (1997b). Here we shall first discuss a numerical algorithm suitable for calculation of inclusion probabilities for a general order sampling scheme and then implement this algorithm for the PPS case. A key ingredient in the method is numerical calculations of the distribution functions of order statistics of a sample of independent, but not necessarily identically, distributed random variables.

Lemma 2. *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 the recursive equation:*

$$F^N_n(t) \quad = \quad F^{N-1}_n(t) + F_N(t)[F^{N-1}_{n-1}(t) - F^{N-1}_n(t)], \tag{2.5}$$

where $F^N_0(t) = 1$, for all N and t.

Proof. Observe that,

$$\{Q^N_{(n)} \leq t\} \Longleftrightarrow \quad \{Q^{N-1}_{(n)} \leq t\} \bigcup (\{Q^{N-1}_{(n-1)} \leq t < Q^{N-1}_{(n)}\} \bigcap \{Q_N \leq t\})$$

so that

$$\begin{aligned} P(Q^N_{(n)} \leq t) \quad &= P(Q^{N-1}_{(n)} \leq t) + (P(Q^{N-1}_{(n-1)} \leq t) \\ &\quad - P(Q^{N-1}_{(n)} \leq t))P(Q_N \leq t), \end{aligned}$$

which is equivalent to (2.5). ■

Returning to the order sampling procedure, the probability of element N belonging to the sample s is

$$\begin{aligned} \pi_N \quad &= \quad P(N \in s) = P(Q^{N-1}_{(n)} > Q_N) \\ &= \quad \int_0^\infty (1 - F^{N-1}_n(t)) f_N(t) dt. \end{aligned} \tag{2.6}$$

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. The algorithm above is related to the one used earlier to calculate the conditional inclusion probabilities under CPS design.

2.3. The Pareto πps Sampling (PPS) case

Consider an Order Sampling procedure as described in Section 2.2 and suppose that $F_i(t) = \theta_i t/(1+\theta_i t)$ is the standard Pareto distribution function with parameter $\theta_i > 0$, for $i = 1, ..., N$. The densities then become $f_i(t) = \theta_i/(1+\theta_i t)^2$, $i = 1, ..., N$. Parallel to this θ parametrisation we shall use an alternative set of parameters which are more directly coupled to the inclusion probabilities: $\lambda_i = F_i(1) = \theta_i/(1+\theta_i)$, $i = 1, ..., N$. This is motivated by the fact that λ_i approximates the inclusion probabilities in case $\sum_U \lambda_i = n$, cf. Rosén (1997b). Let $\tilde{\pi}_i$ denote the inclusion probabilities as functions of λ. The resulting inclusion probabilities $\tilde{\pi}_i$ from the choice λ are not exactly the desired ones, but they are good approximations for large samples.

We have developed a method for calculating the exact inclusion probabilities, given by the integral in (2.6) with help of Lemma 2.2. The exact inclusion probabilities $\tilde{\pi}_i$ are computed by numerical approximations with a computer program. The input of this program is a vector of distribution function values evaluated in $t = 1$, $(\lambda_1, ..., \lambda_N)$. To solve the integral, we first build a recursive function to calculate F_n^N. The function returns for any vector of time points $(t_1, ..., t_k)$ the vector of distribution function values $(F_n^N(t_1), ..., F_n^N(t_k))$. In a second step the integral (2.6) is calculated by using the built-in Matlab function, *quad* which performs numerical integration using the adaptive recursive Simpson's rule, cf. Atkinson (1988). Since the Simpson's algorithm, *quad* in Matlab has too narrow limits for the number of grid points used to give sufficient precision, to calculate integrals of the form

$$\int_0^\infty \mu(t)dt,$$

for different functions μ, we have used the standard algorithm on the rewritten, interval summed, equivalent integral expression

$$\int_0^1 \frac{1}{10} \sum_{i=0}^{9} \left(\mu(i+0.1t) + \mu\left(i + 0.1\frac{1}{t}\right)\frac{1}{t^2} \right) dt.$$

(To avoid division by zero for $t = 0$ we have translated the summand for $i = 0$ with the negligible amount 0.0000000000001). This transformation gives good precision. However the computation gets slower because of the massive amount of summations involved.

Example 3 For the vector $\lambda = (0.1\ 0.2\ 0.3\ 0.5\ 0.9)$, we compute the first order inclusion probabilities, $\tilde{\pi}$, when sampling $n = 2$ elements according to the PPS scheme. The execution time is four minutes and $\sum_{i=1}^N \tilde{\pi}_i = 1.99997669002569$, see Table 2.

Table 2: 1st. order inclusions, PPS.

λ	$\tilde{\pi}$
0.1	0.09455331055179
0.2	0.18973382222461
0.3	0.28982049217919
0.5	0.51794513397826
0.9	0.90792393109184
Sum: 2.0	1.99997669002568

Example 4 Consider the population in Example 2. We compute the first order inclusion probabilities for PPS scheme by using the program in Example 3. The resulting inclusions are equal to the λ-values, with precision 10^{-4}. The execution time is almost seven days in a SGI Origin 2000 computer and the control sum is 79.9994.

3. Inclusion probabilities of second order

An important application of the calculation of the inclusion probabilities of second order occurs in computing the variance for an estimator of a population total by using the Yates-Grundy-Sen variance formula for fixed sample size.

$$V(\hat{t}_y) = -\frac{1}{2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{3.1}$$

Here, π_{ij}, denotes the probability that both elements i and j belong to the sample s. They also derived an estimator of $V(\hat{t}_y)$, cf. Särdal *et. al.* (1992), which is unbiased for a sampling design of fixed size, provided that $\pi_{ij} > 0$ for all $i \neq j \in U$. In the following sections, we derive different methods for numerical calculation of second order inclusion probabilities for both sampling schemes.

3.1. Conditional Poisson Sampling (CPS)

Consider a CPS of size n from $U = \{1, ..., N\}$ with unconditional Bernoulli parameters $p_1, ..., p_N$. We denote by $\breve{\pi}_{ij}$, the second order inclusion probabilities for CPS design,

$$\breve{\pi}_{ij} \quad = \quad P(i, j \in s \mid\mid s \mid = n) = \frac{\sum_{s \in A_n^{ij}} P(s)}{\sum_{s \in A_n} P(s)},$$

where A_n denotes the class of all possible samples of size n and A_n^{ij} is the set of elements of A_n containing both i and j, $i,j = 1,...,N$. 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 and by consideration of the equations,

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

Based on the recursive idea from before a computer program was developed to compute the bivariate inclusion probabilities using this method. But some examples calculated, cf. Aires (1998), showed that this procedure requires more execution time than those calculations made with the method we present below.

Let[2] $\breve{\pi}_{i\setminus j}$ denote the probability that element i but not element j belongs to the sample s, for a Conditional Poisson sample of size n. We denote the odds by $\gamma_i = p_i/(1-p_i)$. From equation (2.1) it may be deduced that,

$$\breve{\pi}_{i\setminus j} = \frac{\sum_{s\in A_n^{i\setminus j}} \frac{p_i}{(1-p_i)} \prod_{k\in s'} p_k/(1-p_k)}{\sum_{s\in A_n} \prod_{l\in s} p_l/(1-p_l)},$$

where $A_n^{i\setminus j}$ is the subset of $s \in A_n$ with $i \in s$, $j \notin s$, and s' is the set s excluding element i. Furthermore,

$$\breve{\pi}_{i\setminus j} = \underbrace{\frac{\sum_{s\in A_n^{j\setminus i}} \frac{p_j}{(1-p_j)} \prod_{k\in s'} p_k/(1-p_k)}{\sum_{s\in A_n} \prod_{l\in s} p_l/(1-p_l)}}_{\breve{\pi}_{j\setminus i}} \frac{\gamma_i}{\gamma_j} = \breve{\pi}_{j\setminus i} \cdot \frac{\gamma_i}{\gamma_j},$$

and by using the odds ratio, we can write,

$$\breve{\pi}_{i\setminus j} = \frac{\gamma_i}{\gamma_j} \cdot \breve{\pi}_{j\setminus i}. \tag{3.2}$$

On the other hand,

$$\begin{cases} \breve{\pi}_i = \breve{\pi}_{i\setminus j} + \breve{\pi}_{ij} \\ \breve{\pi}_j = \breve{\pi}_{j\setminus i} + \breve{\pi}_{ij}. \end{cases} \tag{3.3}$$

By substituting (3.2) in (3.3),

$$\begin{cases} \breve{\pi}_i = \frac{\gamma_i}{\gamma_j} \breve{\pi}_{j\setminus i} + \breve{\pi}_{ij} \\ \breve{\pi}_j = \breve{\pi}_{j\setminus i} + \breve{\pi}_{ij}. \end{cases}$$

[2]This method was developed with Prof. O. Nerman.

Combining the two equations, we get

$$\breve{\pi}_{ij} = \frac{\gamma_i \breve{\pi}_j - \gamma_j \breve{\pi}_i}{\gamma_i - \gamma_j} \tag{3.4}$$

in case $\gamma_i \neq \gamma_j$.

For the case $\gamma_i = \gamma_j$, fix a pair i and j_0 such that $\gamma_i = \gamma_{j_0}$, and observe that $\breve{\pi}_{ij}/\breve{\pi}_i$ may be regarded as the probability of the *j:th* unit in CPS of size $n-1$ taken from $\{1,..,N\} \setminus \{i\}$, so that

$$\frac{\sum_{j:j\neq i} \breve{\pi}_{ij}}{\breve{\pi}_i} = (n-1),$$

and hence

$$(n-1)\breve{\pi}_i = \sum_{j:j\neq i} \breve{\pi}_{ij} = (\sum_{j:\gamma_j \neq \gamma_i} \breve{\pi}_{ij}) + k_i \breve{\pi}_{ij_0},$$

where k_i is the number of elements $j \neq i$ such that $\gamma_j = \gamma_i$ or equivalently $p_j = p_i$. Thus,

$$\breve{\pi}_{ij_0} = \frac{(n-1)\breve{\pi}_i - \sum_{j:\gamma_j \neq \gamma_i} \breve{\pi}_{ij}}{k_i}. \tag{3.5}$$

Note that the number of pairs of different individuals in a sample of size n is always $n \cdot (n-1)/2$ and therefore $\sum_{i<j} \breve{\pi}_{ij} = n \cdot (n-1)/2$ can be used as a check.

Example 5 Second order inclusion probabilities are calculated given $p = (0.1, 0.2, 0.3, 0.5, 0.9)$ and $\breve{\pi} = (0.0695, 0.1543, 0.2600, 0.5732, 0.9431)$. The execution time is 0.0053 seconds and the control sum is 1 with 15 decimals. The results follow in Table 3.

Example 6 For the population in Example 2 and using the same program in Example 5, the second order inclusion probabilities are calculated. The execution time is 7 seconds and the check sum is $\sum_{i<j} \breve{\pi}_{ij} = 3.159999999513763 \cdot 10^3$, which should be compared to $3.16 \cdot 10^3$.

Table 3: $\breve{\pi}_{ij}$, 2nd. order inclusion probabilities, CPS.

		j			
	1	2	3	4	5
1		0.0016	0.0028	0.0065	0.0586
2			0.0063	0.0146	0.1317
i 3				0.0251	0.2258
4					0.5270

3.2. The Pareto πps Sampling (PPS) case

Consider an Order Sampling scheme with N units and sample size n. Then the bivariate inclusion probability of the units $N-1, N$ is given by:

$$\pi_{N-1,N} = P(N-1 \in s, N \in s) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N))$$

$$= \int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1},Q_N)}(t)dt. \tag{3.6}$$

This is a special case of a more general higher order inclusion probability formula derived from the observation that

$$\{N-k+1 \in s, ..., N \in s\} \Longleftrightarrow \quad \{Q_{(n-k+1)}^{N-k} > max(Q_{N-k+1}, ..., Q_N)\},$$

which in general yields

$$P(N-k+1 \in s, ..., N \in s) = P((Q_{(n-k+1)}^{N-k}) > max(Q_{N-k+1}, ..., Q_N))$$

$$= \int_0^\infty (1 - F_{n-k+1}^{N-k}(t)) f_{max(Q_{N-k+1},\cdots,Q_N)}(t)dt.$$

As in the univariate case, the inclusion probability of an arbitrary pair of units $i < j$ may be determined by consideration of the rearranged sequence:

$$Q_1, Q_2, ..., Q_{i-1}, Q_{i+1}, ..., Q_{j-1}, Q_{j+1}, ..., Q_N, Q_i, Q_j.$$

In the Pareto case, by the differentiation chain rule,

$$f_{max(Q_i,Q_j)}(t) = F'_{max(Q_i,Q_j)}$$

$$= \frac{\theta_i}{(1+\theta_i t)^2}\left(1 - \frac{1}{(1+\theta_j t)}\right) + \frac{\theta_j}{(1+\theta_j t)^2}\left(1 - \frac{1}{(1+\theta_i t)}\right).$$

According to the Formula (3.6), a program is built to calculate the second order inclusion probabilities for PPS scheme. Analogous to the one dimensional case, to solve the integral we first build a recursive function to calculate F_{n-1}^{N-2}. In a second step the integral in (3.6) is calculated using the built-in Matlab function *quad* exactly in the same way as the one described in Section 2.3 for the one dimensional inclusions.

Example 7 In this example the input parameters are $\lambda = (0.10.2\ 0.3\ 0.5\ 0.9)$, $N = 5$ and $n = 2$. The resulting bivariate probabilities are shown in Table 4. The control sum is $\sum_{i=1}^{N-1}\sum_{j=1}^{N}\tilde{\pi}_{ij} = 0.99997$. The routines to

Table 4: $\tilde{\pi}_{ij}$, 2nd. order inclusion probabilities, PPS.

		j				
		1	2	3	4	5
	1		0.0033	0.0054	0.0112	0.0747
	2			0.0113	0.0234	0.1517
i	3				0.0375	0.2357
	4					0.4458

calculate second order inclusion probabilities for the Pareto case requires large execution time.

4. Adjusting the parameters

4.1. Conditional Poisson Sampling (CPS) scheme

Assume that the exact inclusion probabilities z_i are given and that we wish to adjust the unconditional inclusion probabilities p_i, so that the conditional inclusion probabilities equal z_i, $i = 1, ..., N$. As shown in Dupačová (1979), the equation system (2.2) has a unique solution, and thus the p_i-values can be found, by solving a non-linear equation system with N unknown variables and $N + 1$ equations:

$$\begin{cases} \dfrac{\sum_{s\in A_n^1}\prod_{j\in s}p_j\ \prod_{k\notin s}(1-p_k)}{\sum_{s\in A_n}\prod_{j\in s}p_j\ \prod_{k\notin s}(1-p_k)} - z_1 = 0 \\ \qquad\qquad\vdots \\ \dfrac{\sum_{s\in A_n^N}\prod_{j\in s}p_j\ \prod_{k\notin s}(1-p_k)}{\sum_{s\in A_n}\prod_{j\in s}p_j\ \prod_{k\notin s}(1-p_k)} - z_N = 0 \\ p_1 + ... + p_N - n = 0. \end{cases} \tag{4.1}$$

There are numerical methods for solving nonlinear equations systems of the form $\mathbf{f}(\mathbf{x}) = 0$ as in (4.1). We choose the built-in function *fsolve* in Matlab, which uses the Gauss-Newton method, cf. Atkinson (1988), to built a program that calculates the vector of adjusted unconditional inclusion probabilities p^A.

Example 8 Given the desired inclusion probabilities z = (0.10.2 0.3 0.5 0.9), we calculate the adjusted unconditional inclusion probability vector p^A. The program is executed a second time with the vector p^A as input to calculate $\pi^A = \breve{\pi}(p_1^A, ..., p_N^A)$. The resulting values π^A are extremely similar to the z values as expected, see Table 5. For larger populations this method

Table 5: Adjusted Conditional Poisson scheme.

z	$\breve{\pi}$	p^A	π^A
0.1	0.0694	0.13283686195059	0.10000001974508
0.2	0.1543	0.23867750414515	0.19999999990091
0.3	0.2600	0.32526121539275	0.30000000063677
0.5	0.5732	0.45945330941274	0.49999999390902
0.9	0.9431	0.84377110396854	0.89999998580823
Sum: 2.0	2.0000	1.99999999486977	2.00000000000001

is very slow. In the case where the population size is $N = 284$ and the sample size is $n = 80$, the execution time to calculate the adjusted inclusion probabilities is 36 hours and the control sum is 80 with precision 10^{-4}.

4.2. Pareto πps Sampling (PPS) scheme

Consider a small Pareto *πps* sample of size n with target inclusion probabilities $z_1, ..., z_N$, strictly positive and less than 1. In this section we introduce a method to adjust the λ-parameters, using the Gauss-Newton procedure to solve the equation system:

$$\begin{cases} \tilde{\pi}_i(\lambda_1, ..., \lambda_N) = z_i, \quad i = 1, ..., N \\ \sum_{i=1}^{N} \lambda_i = n. \end{cases} \tag{4.2}$$

Here $\tilde{\pi}_i(\lambda_1, ..., \lambda_N)$, $i = 1, ..., N$, are derived through the earlier described procedure. The existence and uniqueness of the solution of equation (4.2) has not been justified by a stringent argument but preliminary results indicate that this conjecture is true.

Example 9 Given the vector z of target inclusion probabilities, we calculate the adjusted inclusions for the PPS scheme. The results, denoted by λ^A, are shown in Table 6. We also calculate $\tilde{\pi}^A = \tilde{\pi}(\lambda_1^A, ..., \lambda_N^A)$. Due to the

Table 6: Adjusted inclusion probabilities for PPS.

z	$\tilde{\pi}$	λ^A	$\tilde{\pi}^A$
0.1	0.0946	0.10528414513777	0.09999484693095
0.2	0.1897	0.20958638050326	0.19999469027722
0.3	0.2898	0.30857771776815	0.29999548194969
0.5	0.5179	0.48480955603085	0.49999528497564
0.9	0.9079	0.89174250169828	0.89999591692859
Sum: 2.0	2.0000	2.00000030113832	1.99997622106209

fast convergence for this case, no adjustment routines are needed for larger populations.

4.3. An alternative algorithm

Unfortunately the programs to adjust the parameters in both methods require long execution time if we wish to ensure a high approximation precision. For that reason and lead by observation of the pattern of p^A and λ^A values in relation to the z values, we have also applied a more simple idea to achieve the adjusted values. In the PPS case the alternative adjusting procedure consists in iterating

$$\lambda^A(k+1) = \lambda^A(k) + (z - \tilde{\pi}(\lambda^A(k))) \tag{4.3}$$

until

$$\max \left| \frac{\tilde{\pi}_i(\lambda^A(k))}{z_i} - 1 \right| \leq 10^{-4}, \tag{4.4}$$

given the start value

$$\lambda^A(0) = z. \tag{4.5}$$

After each iteration we adjust the $\tilde{\pi}_i(\lambda^A(k))$ values slightly by normalising,

$$\tilde{\pi}_i(\lambda^A(k)) = \left(\frac{\tilde{\pi}_i(\lambda^A(k))}{\sum_j \tilde{\pi}_j(\lambda^A(k))} \right) \cdot n. \tag{4.6}$$

To get the CPS variant substitute λ by p. This algorithm was implemented for both sampling methods and applied in some of our larger sample size examples. For the population size $N = 284$ and the sample size $n = 80$

for the CPS case, the execution time to calculate the adjusted inclusion probabilities, was 15 minutes and the control sum was 80 with precision 10^{-4}.

5. Conclusions

We have shown that it is feasible to calculate first and second order inclusion probabilities in both CPS and PPS Designs. The program routines provide good numerical precision and reasonable execution times for quite large sample sizes. For the PPS case the routines to calculate the first and second order inclusion probabilities are more time consuming due to the numerical integration required. For the Conditional Poisson case, the execution performance is quite good and the algorithms to compute the inclusion probabilities are fast and effective. Second order inclusion probabilities are as easily calculated as first order ones. Further work to achieve more precision in reasonable execution time in the numerical integration for the Pareto case is in progress, as well as the improvement of the routines that implement the calculations of second order inclusion probabilities for this scheme. It is also feasible to generalise the programs now developed for the PPS procedure to other order sampling schemes with fixed distribution shape.

Acknowledgements. I wish to thank Professor Olle Nerman for his valuable help throughout this work. My sincere thanks also to Björn von Sydow for computational advice.

References

Aires, N. (1998) Exact Inclusion Probabilities for Conditional Poisson Sampling and Pareto πps Sampling Designs, Studies in Applied Probability and Statistics, Mathematical Statistics, Chalmers University of Technology, Göteborg University.

Aires, N. (2000) Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs, *Journal of Statistical Planning and Inference*, **82**, 1-15.

Atkinson, K. (1988). An Introduction to Numerical Analysis, 2nd. ed., Wiley, Singapore.

Dupačová, J. (1979). A note on Rejective Sampling, Contributions to Statistics, Jaroslav Hájek Memorial Volume, Reidel, Holland and Academia, Prague, 71-78.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population, *Annals of Mathematical Statistics*, **35**, 1491-1523.

Hájek, J. (1981) Sampling from a Finite Population, Marcel Dekker, New York.

Ohlsson, E. (1995). Sequential Poisson Sampling, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Report No. 182.

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. (1998). On Inclusion Probabilities of Order Sampling, R&D Report, Research - Methods - Development 1998:2, Statistics Sweden.

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

Paper B

Reprinted from JOURNAL OF STATISTICAL PLANNING AND INFERENCE, Vol. 82, Aires N. Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs, Pages No. 1-15, ©2000, with permission from Elsevier Science.

Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs

Nibia Aires

School of Mathematical and Computing Sciences,
Chalmers University of Technology,
S-412 96 Gothemburg, Sweden

Received 10 June 1999; recieved in revised form 12 October 1999

Abstract

Conditional Poisson Sampling (CPS) and Pareto πps Sampling (PPS) design belong to a class of sampling schemes which have inclusion probabilities proportional to given size measures (πps). Algorithms were introduced to calculate first and second order exact inclusion probabilities for both schemes. In this paper using those algorithms, both methods are compared by calculating Horvitz-Thompson estimators of populations totals and computing their biases and variances. They are also illustrated by examples, and the results in these are compared with known asymptotics.

Keywords: Sampling Theory, Conditional Poisson Sampling, Pareto πps Sampling, Numerical Integration, Algorithms

AMS 1991 subject classification: primary 62D05; secondary 65U05

1. Introduction

Hájek in 1964 showed that conditioning on the sample size n in a Poisson Sampling Design, yields the maximum entropy distribution of the sample s among all sampling procedures of size n, with inclusion probabilities of the individuals equal to those of these Conditional Poisson Sampling (CPS) procedures, see Hájek (1964). In the same paper, he developed an asymptotic theory for inclusion probabilities and estimator distributions in this scheme. Based on a refined asymptotic study of the relation between the conditional and unconditional inclusion probabilities, he also proposed formulas to approximate and adjust the inclusion probabilities, see Hájek (1981).

On the other hand, an asymptotic theory for the Pareto πps Sampling (PPS) design is introduced by Rosén (1997a, 1997b). He shows that this scheme is asymptotically uniformly optimal among the procedures which have inclusion probabilities proportional to given size measures (πps) and belong to a class of sampling schemes called Order Sampling with fixed distribution shape. In addition, the asymptotic results of previous studies make us believe that CPS and PPS schemes are very similar.

In a recent paper, Aires (1999), algorithms are introduced to calculate first and second order exact inclusion probabilities for both schemes. This fact make it possible to compare the CPS design with the PPS scheme. In Aires (1999) very small illustrating examples are given. Here we shall study the properties of the two schemes more systematically and compare convergence behaviour of first and second order inclusions and estimator distributions, using the results in Aires (1999) as our main tool.

In the first part of this paper, we present calculations of first and second order exact inclusion probabilities for both sampling schemes for small and moderate population sizes. Comparisons between methods are carried out by calculations of the relative biases and variances of the estimators of the population totals. The results show that for the PPS design the inclusion probabilities converge faster to the given inclusion probabilities than for the CPS scheme. Additionally the bound of the measure of variance dissimilarities and the bound for the relative biases are calculated. The results obtained in those corroborate the similarity of both schemes. Finally examples for larger population sizes are presented. In most given examples it is observed that even if the asymptotic variances are very close in both schemes, the asymptotic variance for PPS design approximates the real variance better. This real variance is slightly smaller for CPS.

2. Conditional Poisson and Pareto πps Sampling Designs

Poisson Sampling is a method for choosing a sample s, of random size $|\,s\,|$, from a finite population U consisting of N individuals. Each individual i in the population has a predetermined probability p_i of being included in the sample s. A Poisson sample may be realised by using N independent Bernoulli trials to determine whether the individual under consideration is to be included in the sample s or not, i.e. by picking the individuals independently of each other. CPS design may be defined as Poisson Sampling conditioned by the requirement that the sample has fixed size n, cf. Hájek (1964). This sampling procedure can be realised by a rejection procedure where independent samples are sequentially drawn until we get a sample with the required size n. Let A_n denote the class of all possible samples of size n and A_n^i the set of elements of A_n containing i, so that

$$\breve{\pi}_i \quad = \quad \breve{\pi}_i(p) = P(i \in s \mid |\,s\,| = n) = \frac{\sum_{s \in A_n^i} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)}{\sum_{s \in A_n} \prod_{j \in s} p_j \prod_{k \notin s}(1-p_k)},$$

$i = 1, ..., N$, are the inclusion probabilities of the individuals in the CPS procedure. These inclusions easily can be calculated using the recursive procedure in Aires (1999). Second order inclusion probabilities may be calculated in a similar way, but we choose to use the alternative formulas,

$$\breve{\pi}_{ij} \quad = \quad \frac{\gamma_i \breve{\pi}_j - \gamma_j \breve{\pi}_i}{\gamma_i - \gamma_j}$$

where $\gamma_i = p_i/(1-p_i)$ valid in the case $\gamma_i \neq \gamma_j$, and

$$\breve{\pi}_{ij_0} \quad = \quad \frac{(n-1)\breve{\pi}_i - \sum_{j:\gamma_j \neq \gamma_i} \breve{\pi}_{ij}}{k_i}$$

for the case $\gamma_i = \gamma_{j_0}$, k_i being the number of elements $j \neq i$ such that $\gamma_j = \gamma_i$ or equivalently $p_j = p_i$, see Aires (1999).

An Order Sampling scheme is originally defined by considering independent ranking random variables $Q_1, Q_2, ..., Q_N$ with distribution functions $F_i(t)$ and Lebesgue density $f_i(t)$, $0 \leq t < \infty$, $i = 1, ..., N$. The units with the n smallest Q-values constitute the sample. An important subclass is derived by the assumption that all F_i belong to the same scale family and in the case these distribution functions are defined as $F_i(t) = \theta_i t/(1+\theta_i t)$, $t > 0$, $\theta_i > 0$, we have a PPS procedure. Parallel to this θ parametrisation we shall use an alternative set of parameters which are more directly coupled to the inclusion probabilities: $\lambda_i = F_i(1) = \theta_i/(1+\theta_i)$, $i = 1, ..., N$. This

is motivated by the fact that λ_i approximates the inclusion probabilities in case $\sum_U \lambda_i = n$, cf. Rosén (1997b).

The probability of element N belonging to the sample s is

$$\begin{aligned} \tilde{\pi}_N(\lambda) &= P(N \in s) = P(Q_{(n)}^{N-1} > Q_N) \\ &= \int_0^\infty (1 - F_n^{N-1}(t)) f_N(t) dt. \end{aligned} \tag{2.1}$$

Here, $F_n^{N-1}(t)$ is the distribution function of $Q_{(n)}^{N-1}$, the *n:th* order statistic among $Q_1, Q_2, ..., Q_{N-1}$. The inclusion probability of any other unit i, $\tilde{\pi}_i$, 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. It is straightforward to generalise this procedure to higher order inclusion probabilities, cf. Aires (1999). Here we shall use the second order inclusion probability formula

$$\begin{aligned} \pi_{N-1,N} &= P(N-1 \in s, N \in s) = P(Q_{(n-1)}^{N-2} > max(Q_{N-1}, Q_N)) \\ &= \int_0^\infty (1 - F_{n-1}^{N-2}(t)) f_{max(Q_{N-1}, Q_N)}(t) dt, \end{aligned} \tag{2.2}$$

for units $N-1, N$. The functions $F_n^{N-1}(t)$ and $F_{n-1}^{N-2}(t)$, in (2.1) and (2.2), respectively, are calculated by using a recursive procedure and then the integrals are computed using numerical methods, cf. Aires (1999).

2.1. Background

In many situations it is of interest to estimate the population total of some study variable y known only for a sample s from the population $U = \{1, ..., N\}$. The estimation of the population total may be improved in cases where for each individual i in the population a positive quantity x_i is known and it is believed that x_i is proportional to a study variable y_i. In order to use this information to get good estimators of the population total, the choice of the sample s can be made using inclusion probabilities π_i proportional to x_i and then the unbiased Horvitz-Thompson estimator, $\sum_{i \in s} y_i/\pi_i$, may be adopted to estimate the population total. In some cases, it may not be possible to select a sample based on strictly proportional inclusion probabilities π_i, simply by taking, $\pi_i = nx_i/\sum_U x_j$, $i = 1, ..., N$ since one or more of the $\pi's$ may exceed 1. This obstacle is usually circumvented in practice by introducing a "take all" stratum of units with the

largest sizes. We shall assume from now on that $nx_i < \sum_U x_j$, $i = 1, ..., N$. Second order inclusion probabilities π_{ij}, are of great interest to establish the variance, by using, for instance, the Yates, Grundy and Sen formula for fix size sampling designs:

$$-\frac{1}{2}\sum_{i,j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2. \tag{2.3}$$

This formula in its turn can be used to derive unbiased variance estimators. The theory above is developed for the situation where the vector π is known exactly. Next we shall consider a situation where z is the vector of ideal inclusion probabilities and then study several CPS and PPS schemes in this situation. Since the results in Aires (1999) made it possible to exhibit manageable exact expressions for first and second order inclusion probabilities for both schemes, standard results can be applied for the linear estimators and comparisons can be made between methods as well as to earlier asymptotic results. First order exact inclusion probabilities for CPS and PPS designs can be calculated based on these Bernoulli parameters and by using the algorithms in Aires (1999). We shall refer to these procedures as the straightforward calculations when we use $p = z$ for Conditional Poisson and $\lambda = z$ for the Pareto case. The results are denoted by π^R and $\tilde{\pi}^R$ respectively, and $\sum \pi_i^R = n$ and $\sum \tilde{\pi}_i^R = n$ can be used as checks. In the earlier work we also develop algorithms to adjust the p-parameters in the CPS case and the λ-parameters in the PPS scheme to achieve any desired inclusion probability vector z exactly. We refer to these procedures as the adjusted procedures and use a superscript A in the notation, i.e. p^A and λ^A denote the parameter vectors that give exact inclusions z. Analogously, second order exact inclusion probabilities for both methods may be calculated using algorithms in Aires (1999). The bivariate inclusion probabilities are denoted by π_{ij}^R and $\tilde{\pi}_{ij}^R$, for the unadjusted CPS and for the unadjusted PPS case, respectively. Similarly, for the adjusted procedures, second order inclusion probabilities are denoted by π_{ij}^A and $\tilde{\pi}_{ij}^A$ respectively. In addition, as the number of different pairs of individuals in a sample of size n is always $n(n-1)/2$, the sum, $\sum_{i<j}\pi_{ij} = n\cdot(n-1)/2$ can be used as a check.

In the sequel we shall always start with an ideal inclusion probability vector z and let the estimator of the population total be

$$\hat{t}_y = \sum_s y_i/z_i \tag{2.4}$$

irrespective of which of the four distributions that are used. Moreover, the variance is given by

$$V(\hat{t}) = -\frac{1}{2}\sum_{i,j\in U}(\pi_{ij}-\pi_i\pi_j)\left(\frac{y_i}{z_i}-\frac{y_j}{z_j}\right)^2, \tag{2.5}$$

where π_{ij} and π_i are the exact inclusions of each method, see Särndal *et. al* (1992). Equally important, when comparing sampling methods, is the calculation of the theoretical bounds for the relative bias and variance properties. For the former, Rosén (1998) develops a goodness measure comparing inclusion probabilities and their approximations under the PPS scheme. He shows that for a study variable $y > 0$, the relative absolute bias $\tilde{B}$ for the estimator $\hat{t}_y$, using the parameters $\lambda = z$, can be bounded by

$$\tilde{B} = \left| \frac{\tilde{E}(\hat{t}_y) - t_y}{t_y} \right| = \left| \frac{1}{t_y} \sum y_i (\frac{\tilde{\pi}_i}{\lambda_i} - 1) \right| \leq max \left(\left| \frac{\tilde{\pi}_i^R}{z_i} - 1 \right| \right) = \tilde{B}_U.$$

In the same way, the bound for the relative bias B can be derived for the CPS case, $B \leq max \left| (\pi_i^R / z_i) - 1 \right| = B_U$.

Using a similar idea, a bound of variance dissimilarities of the adjusted procedures, may be derived by first defining,

$$\psi = \frac{\mid \tilde{V}^A(\hat{t}_y) - V^A(\hat{t}_y) \mid}{V^A(\hat{t}_y)} =$$

$$\frac{\left| -\frac{1}{2} \sum_{i,j \in U} (\tilde{\pi}_{ij}^A - z_i z_j)(\frac{y_i}{z_i} - \frac{y_j}{z_j})^2 + \frac{1}{2} \sum_{i,j \in U} (\pi_{ij}^A - z_i z_j)(\frac{y_i}{z_i} - \frac{y_j}{z_j})^2 \right|}{-\frac{1}{2} \sum_{i,j \in U} (\pi_{ij}^A - z_i z_j)(\frac{y_i}{z_i} - \frac{y_j}{z_j})^2}$$

where $\tilde{V}^A$ and V^A denote the adjusted variance for the PPS and the CPS cases respectively. Next, by some straightforward algebra, and using that $\pi_{ij}^A \leq z_i z_j$ we can bound this expression by

$$max_{i \neq j} \frac{\mid \tilde{\pi}_{ij}^A - \pi_{ij}^A \mid}{\mid \pi_{ij}^A - z_i z_j \mid} = \psi_U. \tag{2.6}$$

While B_U and $\tilde{B}_U$ can be seen to be sharp bounds, ψ_U is typically an overestimate of the possible values of ψ for varying y's. In the case of variance comparisons we do two things, we compare the real variances for the adjusted methods and on the other hand, we compare each of the asymptotic variances with the corresponding exact variance.

3. The implementation

A computer system was developed to make comparisons between CPS and PPS designs in terms of estimators of population totals, biases and variances.

3.1. The input parameters

The program gives the option to load a data file consisting of a column with the study variable y and a column with an auxiliary variable x. For simulation convenience we load the MUCLUS file, which is the clustered MU284 population readily available from Statistics Sweden, at http://lib.stat.cmu.edu/datasets/mu284, to illustrate results for small and moderate population sizes. The MUCLUS population file consists of information about 284 Swedish municipalities clustered in 50 clusters. The variable P85, the municipalities population in 1985 (in thousands) will be used as the study variable y and P75, the municipalities population in 1975 (in thousands) will be used as the auxiliary variable x which we believe is proportional to y =P85. The user is prompted to enter the size of the population N, $N < 50$ and the sample size $n < N$. The population will consist of the first N observations in the MUCLUS file. Based on this information a vector z of desired proportional inclusion probabilities satisfying that $\sum_{i=1}^{N} z_i = n$ is created from P75.[3] For examples with larger population sizes we load the MU284 file, which consists of all the 284 Swedish municipalities. We adopt again y = P85 and x = P75 as the study and the auxiliary variable respectively. For other examples we choose to work with the 271 smallest municipalities of the MU284 file.

3.2. Calculations for Conditional Poisson Sampling (CPS) and the Pareto πps Sampling (PPS) case

1. Given that $p = z$, the program calculates first order inclusion probabilities for the CPS scheme, π^R. For $\lambda = z$, the program calculates first order inclusion probabilities for the PPS scheme, $\tilde{\pi}^R$.

2. The program also computes the adjusted inclusion probabilities p^A, pertaining to z, for CPS and the adjusted inclusion probabilities λ^A, for PPS.

3. Using the vector p^A as input, first order inclusion probabilities are calculated for CPS, and the resulting vector of inclusions is denoted by π^A. For PPS, using the vector λ^A as input, first order inclusion probabilities are calculated and the resulting values are denoted by $\tilde{\pi}^A$. Observe that π^A and λ^A should ideally equal z.

4. The second order inclusion probabilities for the straightforward schemes, are calculated using the input vectors z and π^R for CPS, and z and $\tilde{\pi}^R$

[3]The programs are written in Matlab (ver. 5.2) for the Unix system, however some routines are rewritten in Fortran77 to obtain better performance.

for PPS. The resulting bivariate inclusion probabilities are denoted by π_{ij}^R and by $\tilde{\pi}_{ij}^R$ for CPS and PPS respectively.

5. The second order inclusion probabilities for the adjusted CPS scheme are calculated with input vectors p^A and π^A, and the resulting values are denoted by π_{ij}^A. Analogous for PPS, the adjusted values $\tilde{\pi}_{ij}^A$ are calculated with input vector λ^A and $\tilde{\pi}^A$.

6. In the case we select to work with the study variable y, the program determines the relative absolute bias B for the population total estimator $\hat{t}_y$, in the straightforward CPS case and the relative absolute bias $\tilde{B}$ for the straightforward PPS.

7. The program also calculates the real variance given by (2.3) for both sampling methods. Moreover, for the adjusted CPS scheme, the asymptotic variance, cf. Hájek (1964), given by

$$AV^A = \frac{N}{N-1}\sum_{i=1}^{N}\left(\frac{y_i}{z_i} - \frac{\sum_j y_j \frac{p_j^A}{z_j}(1-p_j^A)}{\sum_j p_j^A(1-p_j^A)}\right)^2 \cdot p_i^A(1-p_i^A).$$

is calculated. Similarly, for the adjusted PPS scheme, the asymptotic variance, cf. Rosén (1997b), given by

$$AV^A = \frac{N}{N-1}\sum_{i=1}^{N}\left(\frac{y_i}{z_i} - \frac{\sum_j y_j \frac{\lambda_j^A}{z_j}(1-\lambda_j^A)}{\sum_j \lambda_j^A(1-\lambda_j^A)}\right)^2 \cdot \lambda_i^A(1-\lambda_i^A) \tag{3.1}$$

is also calculated. The factor $N/(N-1)$, is motivated by an analogous ad hoc argument as that given by Rosén (1997b) for the Pareto case, and it is used here to make comparisons between the methods easier.

8. The program calculates the bound for the relative absolute bias, B_U, $\tilde{B}_U$, respectively for CPS and PPS.

The variance dissimilarity ψ and its bound ψ_U are also calculated.

4. Comparisons between methods

The objective of this paper is to compare the CPS and the PPS designs. For that purpose the following section presents evaluations and results of comparisons for small, moderate and larger population sizes.

For all cases presented below first and second order inclusion probabilities are calculated for the straightforward and the adjusted procedures for both methods.

In the next examples, we use the MUCLUS file to illustrate the convergence behaviour of first order inclusion probabilities for small and moderate sample sizes, as well as the behaviour of biases and estimator variances. Then, the aim is to estimate the population total of the study variable $y = \text{P85}$, using inclusion probabilities proportional to the auxiliary variable $x = \text{P75}$, see Section 3.1 for more details. For these cases tables containing first order inclusion probabilities for the straightforward and adjusted procedures for both methods are presented. Second order inclusion probabilities are summarised by tables showing results of calculations of the estimator variances and the dissimilarity bounds for these adjusted schemes.

For the examples of larger population sizes the file MU284 is used, see Section 3.1, where again, two variables are recorded for each element in the file: the population of Swedish municipalities in 1985, $y = \text{P85}$, and the population of Swedish municipalities in 1975, $x = \text{P75}$. For these cases tables for first order inclusion probabilities are omitted, but tables of the estimator variances and their bounds for the adjusted schemes are presented.

The bias and the bound of the bias for the straightforward procedures are also calculated for all examples.

5. Bias and variances

In the following examples we will use the option to include in the calculations the study variable $y = \text{P85}$. In short, the program works with the MUCLUS data file which consists of information on $N = 50$ clustered Swedish municipalities. For each example we choose the population size and the sample size. Then, the inclusion probabilities are determined proportional to the auxiliary variable $x = \text{P75}$.

5.1. Example $N = 5$, $n = 2$

In the first example we choose $N = 5$ and $n = 2$ and we compute first and second order inclusion probabilities. From the results shown in Table 1, we already note that the approximation of first order inclusion probabilities in the PPS case to the given z-values is better than in the CPS case. For the PPS case B_U, B should be read as $\tilde{B}_U$, $\tilde{B}$. Table 2 describes the relative absolute biases and variances. The former, as well as its bound, is smaller for the PPS case. The variances are very similar but, on the other hand, the asymptotic variance is better for the CPS case.

Table 1: Exact and adjusted inclusion probabilities, *N=5* and *n=2*.

y	z	π^R	p^A	π^A	$\tilde{\pi}^R$	λ^A	$\tilde{\pi}^A$
28	0.3218	0.3031	0.3371	0.3219	0.3187	0.3248	0.3218
36	0.3448	0.3303	0.3562	0.3448	0.3423	0.3472	0.3448
31	0.3678	0.3583	0.3750	0.3678	0.3661	0.3694	0.3678
39	0.4598	0.4750	0.4478	0.4598	0.4623	0.4573	0.4598
43	0.5057	0.5333	0.4839	0.5057	0.5105	0.5012	0.5058
Sum: 177	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000

Table 2: *N=5* and *n=2*, z as in Table 1.

Method	B_U	B	V^A	AV^A
CPS	0.05816923	0.00177287	84.22	85.11
PPS	0.00972744	0.00030478	84.21	85.94
Variance comparison:	$\psi_U = 0.00665939$	$\psi = 0.00006109$		

5.2. Example $N = 14$, $n = 5$

We now consider the population described above with $N = 14$ elements and the sample size $n = 5$. Table 3 shows the calculations of exact and adjusted inclusion probabilities for the CPS and the PPS schemes. We see that the first order inclusion probabilities for the PPS scheme are very close to the intended z-values. As it is shown in Table 4, the relative bias under PPS design, as well as its bound, is smaller than for the CPS case. The asymptotic variance approximates the variance better for the PPS case but the real variance is slightly smaller for CPS design. The values of the second order inclusion probabilities are very similar, this is observed from the variance comparison in Table 4. The control sum of second order inclusions is 10.0000 for the CPS and 9.99999999999996 for the PPS case.

5.3. Example $N = 50$, $n = 5$ and $n = 9$

In the next two examples we consider the cases where $N = 50$, $n = 5$ and $N = 50$ and $n = 9$. We omit here the results for first order inclusion probabilities for the straightforward and adjusted schemes, but we note that the $\tilde{\pi}^R$-values are very similar to the z-values under the PPS scheme. The smallest z_i value is 0.0171 and the largest is 0.54143 in case $n = 5$. Similarly, for $n = 9$ the smallest is 0.03079 and the largest is 0.9746.

Table 3: Exact and adjusted inclusion probabilities, *N=14* and *n=5*.

y	z	π^R	p^A	π^A	$\tilde{\pi}^R$	λ^A	$\tilde{\pi}^A$
28	0.2035	0.1946	0.2121	0.2035	0.2030	0.2040	0.2035
36	0.2180	0.2094	0.2263	0.2180	0.2175	0.2185	0.2180
31	0.2326	0.2244	0.2404	0.2326	0.2321	0.2331	0.2326
39	0.2907	0.2851	0.2959	0.2907	0.2903	0.2911	0.2907
43	0.3198	0.3160	0.3232	0.3198	0.3195	0.3201	0.3198
53	0.3706	0.3707	0.3705	0.3706	0.3706	0.3707	0.3706
50	0.3779	0.3786	0.3772	0.3779	0.3779	0.3779	0.3779
55	0.3924	0.3943	0.3906	0.3924	0.3925	0.3924	0.3924
54	0.4070	0.4101	0.4040	0.4070	0.4072	0.4068	0.4070
54	0.4070	0.4101	0.4040	0.4070	0.4072	0.4068	0.4070
57	0.4142	0.4180	0.4107	0.4142	0.4145	0.4140	0.4142
59	0.4215	0.4259	0.4173	0.4215	0.4218	0.4212	0.4215
62	0.4506	0.4576	0.4440	0.4506	0.4511	0.4501	0.4506
66	0.4942	0.5051	0.4840	0.4942	0.4950	0.4934	0.4942
Sum: 687	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000	5.0000

Table 4: *N=14* and *n=5*, z as in Table 3.

Method	B_U	B	V^A	AV^A
CPS	0.04368502	0.00033962	179.5383	185.7819
PPS	0.00246133	0.00002088	179.5504	182.2331
Variance comparison:	$\psi_U = 0.00333009$	$\psi = 0.00006716$		

From Tables 5 and 6, we observe that the relative bias is small for both schemes but slightly smaller under the Pareto case. The real variances are very similar for both sampling schemes. In addition, the asymptotic variance, for the respective schemes, approximates the real variance better as the sample size increases. This approximation is better under PPS design.

6. Examples with larger populations

From the MU284 population we choose now the 271 smallest communities to be able to vary the sample size and still have the z-values proportional to x = P75. For this population, where $N = 271$, calculations of first and second order inclusion probabilities are realized for different sample sizes, as for the other examples. Results for measures of variance dissimilarities

Table 5: *N=50* and *n=5*.

Method	B_U	B	V^A	AV^A
CPS	0.04589144	0.00047470	39098.17	39351.00
PPS	0.00258615	0.00002201	39088.52	38963.56
Variance comparison:	$\psi_U = 0.01593345$	$\psi = 0.00024693$		

Table 6: *N=50* and *n=9*.

Method	B_U	B	V^A	AV^A
CPS	0.02911717	0.00023734	18849.45	18937.27
PPS	0.00095265	0.00000803	18872.96	18795.31
Variance comparison:	$\psi_U = 0.48211450$	$\psi = 0.0012472$		

and the variance bound, as well as bounds for the relative bias for the straightforward procedures are shown in Table 7 and in Figures 1 to 4.

Table 7: *N=271*.

n	ψ	ψ_U	B	B_U	$\tilde{B}$	$\tilde{B}_U$
5	0.000044	0.000057	0.000029	0.008773	0.00000026	0.000110
12	0.000120	0.001672	0.000028	0.008211	0.00000025	0.000097
20	0.000199	0.003187	0.000027	0.007487	0.00000023	0.000081
30	0.000286	0.005580	0.000025	0.006530	0.00000021	0.000061
50	0.000379	0.013475	0.000021	0.005838	0.00000015	0.000027
60	0.000373	0.023121	0.000017	0.005882	0.00000011	0.000021
65	0.000359	0.066843	0.000015	0.005876	0.00000009	0.000019

We observe that the measure for the variance dissimilarities is very small even for larger sample sizes. This value, as well as its bound increases as the sample size increases. On the other hand, the bound for the relative bias decreases as the sample size n increases for both sampling schemes. But for the PPS case, these values are much smaller than for the CPS case.

In order to appreciate the behaviour of these measures as the sample size increases the plots are presented in different scales.

In Table 8 results for variances calculations are shown. The real variance and the asymptotic variance are denoted by V^A, AV^A and $\tilde{V}^A$, $A\tilde{V}^A$ for CPS and for PPS schemes respectively. We note that the real variance for both methods are very similar. The approximation of the asymptotic

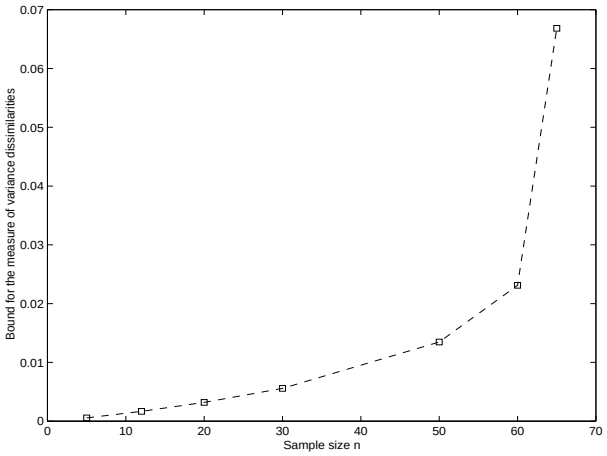

Figure 1: Bound for the Variance Measure of Dissimilarities, $N = 271$.

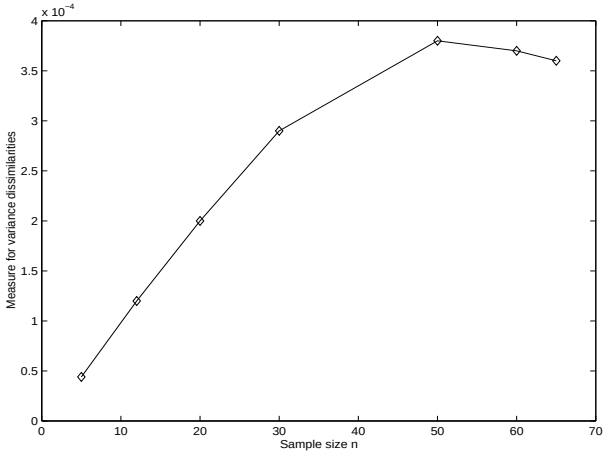

Figure 2: Variance Measure of Dissimilarities, $N = 271$.

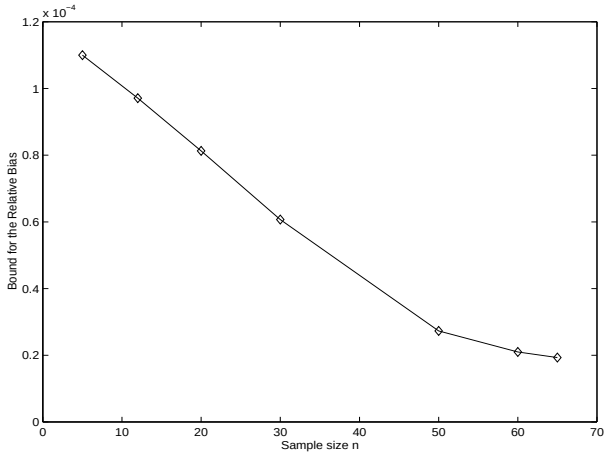

Figure 3: Bound for the Relative Bias, Pareto πps, $N = 271$.

variance to the real variance is better for the PPS case, but the convergence is better for both schemes, as the sample size increases.

The computer programs to make this calculations work very fast since these are developed in Fortran 90 and Fortran 77. For small and moderate populations all results are calculated in some minutes. For larger populations and larger sample sizes the execution time can take approximately one hour. There is a big difference between the complexity of the procedures especially the second order inclusions in the Pareto case take 50 minutes for a population of size $N = 271$ and a sample of size $n = 65$ when for the CPS case, second order inclusion probabilities take the same time for a population of size $N = 542$ and a sample of size $n = 120$.

7. Overall conclusions

7.1. Comparisons of the exact and approximative methods

The implementations made it possible to compare the two methods with each other and to study earlier asymptotically motivated approximations. This has been done with help of a specially designed computer program. The most important consequences of these investigations are sketched next.

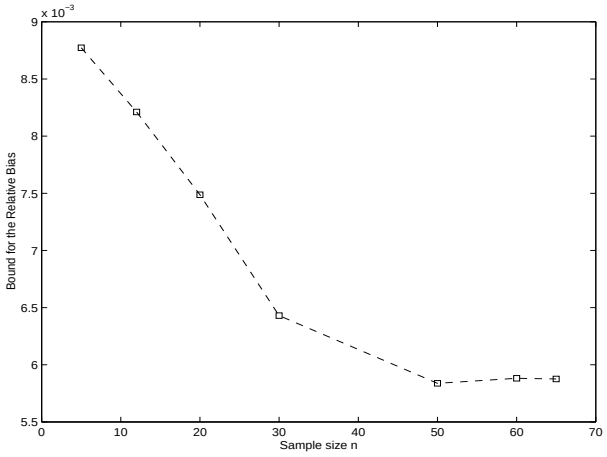

Figure 4: Bound for the Relative Bias, Conditional Poisson, $N = 271$.

There are essentially four main alternatives (described in this paper) for an investigator who is going to make a πps sample.

- CPS with straightforward approximations of inclusions and variances.
- CPS with adjusted unconditional inclusions and corresponding exact inclusions of second order.
- PPS with straightforward approximations of inclusions and variances.
- PPS with adjusted parameters and corresponding exact inclusions of second order.

Both of the approximative methods are easy to implement, but the straightforward PPS method works much better than the straightforward CPS method. On the other hand if one chooses between the more demanding unbiased methods, the resulting variances and second order inclusion probabilities are very similar. Still, there may be advantages with the PPS in connection with multiple or "permanent" samples, where one wants big overlaps between the samples, cf. Ohlsson (1995).

Two natural alternative "exact procedures" that are simpler to implement and faster than the adjustment procedures are derived by the use of π^R, π_{ij}^R, $\tilde{\pi}^R$, $\tilde{\pi}_{ij}^R$, a.s.f. in construction of the Horvitz-Thompson estimators

Table 8: *N=271.*

n	V^A	AV^A	$\check{V}^A$	$A\check{V}^A$
5	56186.76	56232.46	56189.26	56146.85
12	22665.37	22686.72	22668.08	22650.90
20	13087.68	13101.98	13090.29	13080.40
30	8298.67	8309.31	8301.03	8294.88
50	4467.02	4474.30	4468.71	4465.76
60	3508.87	3514.99	3510.18	3508.08
65	3140.27	3145.85	3141.40	3139.61

and in the calculation of variances. These latter alternatives have not been systematically studied in this paper, but we believe that they, in typical applications, will have properties similar to the two studied.

Another conclusion from the results is that the convergence behaviour of first order inclusion probabilities is very fast for the PPS method. This approximation is good also for very small population sizes. We also observe that the PPS method has less bias in all cases and the approximation of the asymptotic variance to the real variance is better for this design. At the same time, this approximation is better for both schemes as the sample size increases.

Of course, this is a first approach to study and compare the characteristics of these two methods. To really understand for example the asymptotic behaviour larger and more systematic studies of the relation between the p, λ and π parameters are needed. A such study for the PPS is under development (Aires & Rosén report).

Acknowledgements

I wish to thank Professor Olle Nerman for his valuable help throughout this work.

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*, **4**, pp. 457-469.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying

Probabilities from a Finite Population. *Annals of Mathematical Statistics.* **35**, pp. 1491-1523.

Hájek, J. (1981). *Sampling from a Finite Population.* Marcel Dekker, New York.

Ohlsson, E. (1995). Sequential Poisson Sampling. Report No. 182. Institute of Actuarial Mathematics and Mathematical Statistics. Stockholm University, Sweden.

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

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

Rosén, B. (1998). On Inclusion Probabilities of Order Sampling. R&D Report, Research - Methods - Development No. 2, Statistics Sweden.

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

Paper C

Inclusion Probabilities of Higher Order for Conditional Poisson Sampling

Nibia Aires

School of Mathematical and Computing Sciences,
Chalmers University of Technology,
S-412 96 Gothemburg, Sweden

24 November 1999

Abstract

Conditional Poisson Sampling design as developed by Hajék may be defined as a Poisson sampling conditioned by the requirement that the sample has fixed size. Algorithms were implemented to calculate the first and second order conditional inclusion probabilities given the inclusion probabilities under Poisson Sampling. Numerical methods were also introduced to compute the unconditional inclusion probabilities when the conditional inclusion probabilities are predetermined. In this paper we extend the second order algorithms to a recursive fast procedure to derive higher order inclusion probabilities for Conditional Poisson Sampling design.

Keywords: Sampling Theory, Conditional Poisson Sampling, Algorithms

AMS 1991 subject classification: 62D05, 65U05, 62-04

1. Introduction

Poisson Sampling is a method for choosing a sample s, of random size $|s|$, from a finite population U consisting of N individuals. Each individual i in the population has a predetermined probability p_i of being included in the sample s. A Poisson sample may be realised by using N independent Bernoulli trials to determine whether the individual under consideration is to be included in the sample s or not. Hájek (1964) showed that conditioning on the sample size n in a Poisson Sampling design, yields the maximum entropy distribution of s among all sampling procedures of size n, with inclusion probabilities of the individuals equal to those of this Conditional Poisson Sampling (CPS) procedure. In fact, the N trials form one experiment. Any experiment that results in other than n out of the N individuals

being picked is rejected. One performs sequentially independent experiments until one of the experiments results in n out of the N individuals being picked. Let A_n denote the class of all possible samples of size n and A_n^i the set of elements of A_n containing i, so that

$$\breve{\pi}_i = \breve{\pi}_i(p) = P(i \in s \mid |s| = n) = \frac{\sum_{s \in A_n^i} \prod_{j \in s} p_j \prod_{k \notin s}(1 - p_k)}{\sum_{s \in A_n} \prod_{j \in s} p_j \prod_{k \notin s}(1 - p_k)}, \tag{1.1}$$

$i = 1, ..., N$, are the inclusion probabilities of the individuals in the CPS procedure. It is hardly ever true that $\breve{\pi}_i = p_i$. Nevertheless a choice of the p_i's can be made by solving the following equation system, for any given set of probabilities z_i, $i = 1, ..., N$ satisfying $\sum_{i=1}^N z_i = n$,

$$\breve{\pi}_i(p) = z_i, \ i = 1, ..., N. \tag{1.2}$$

Furthermore, the equation system (1.2) has a unique solution such that $\sum_{i=1}^N p_i = n$, as shown by Dupačová (1979). Hájek also showed that for large samples one can let p_i equal the desired conditional inclusion probabilities z_i to get a good approximation of the solution of equation (1.2). For moderate sample sizes he also suggested more elaborate approximative formulas in order to get more precise inclusion probabilities or vice versa approximative adjustment of the unconditional inclusions. In Aires (1999), methods are given to calculate the exact inclusion probabilities of first and second order. In the same paper, numerical methods are also given to solve (1.2). The second order conditional inclusions, $\breve{\pi}_{ij}$, turned out to have a surprisingly simple relation to the first order conditional inclusions and the unconditional inclusion probabilities. Namely,

$$\breve{\pi}_{ij} = \frac{\gamma_i \breve{\pi}_j - \gamma_j \breve{\pi}_i}{\gamma_i - \gamma_j} \tag{1.3}$$

in case $\gamma_i \neq \gamma_j$, where $\gamma_i = p_i/(1 - p_i)$. For the case $\gamma_i = \gamma_j$, consider j_0 such that $\gamma_i = \gamma_{j_0}$, then

$$\breve{\pi}_{ij_0} = \frac{(n-1)\breve{\pi}_i - \sum_{j:\gamma_j \neq \gamma_i} \breve{\pi}_{ij}}{k_i}, \tag{1.4}$$

where k_i is the number of elements j such that $j \neq i$ and $\gamma_j = \gamma_i$. A straightforward, but more time consuming, generalisation of the numerical procedures for calculations of first order inclusions were also given for second order inclusions.

In this paper we generalise these two methods to calculate inclusion probabilities of higher order for the CPS design. These inclusion probabilities are of interest for calculation and estimation of moments of higher order, e.g. to facilitate corrections on the coverage accuracy of the confidence interval of the estimator, cf. Thompson (1997) and Hall (1992).

2. Inclusion probabilities of higher order

For simplicity of presentation and ease of understanding, the results will be explicitly given, but the argument will be presented only for third order inclusion probabilities. The generalisation to inclusion probabilities of higher order should be straightforward once the principal ideas have been grasped.

Consider a CPS of size n from $U = \{1, ..., N\}$ with unconditional Bernoulli parameters $p_1, ..., p_N$. We denote by $\breve{\pi}_{ijk}$, the inclusion probabilities of third order, $n \geq 3$ and by $P(s)$ the probability distribution for the unconditional Poisson design. Moreover

$$\breve{\pi}_{ijk} = P(i, j, k \in s \mid \mid s \mid = n) = \frac{\sum_{s \in A_n^{ijk}} P(s)}{\sum_{s \in A_n} P(s)},$$

where A_n denotes the class of all possible samples of size n and A_n^{ijk} is the set of elements of A_n containing elements i, j and k; $i, j, k = 1, ..., N$. The inclusion probability of third order of units i, j, k to be included in the sample s, $i < j < k$, can be derived similarly as in the univariate case, using Lemma 1 in Aires (1999). In that paper, it is shown that the quantities,

$$S_n^N(p_1, ..., p_N) = \sum_{s \in A_n(N)} \prod_{i \in s} p_i \prod_{j \notin s} (1 - p_j) \tag{2.1}$$

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

$$S_0^N = (1 - p_1)(1 - p_2) \cdots (1 - p_N)$$

and

$$S_N^N = p_1 p_2 \cdots p_N.$$

Thus the third order inclusion probabilities are obtained by consideration of the equations,

$$\breve{\pi}_{ijk} = \frac{p_i p_j p_k S_{n-3}^{N-3}([p_i, p_j, p_k]^c)}{S_n^N(p_1, ..., p_N)},$$

where $[p_i, p_j, p_k]^c$ denotes the vector of inclusion probabilities $[p_1, ..., p_N]$ excluding elements p_i, p_j, and p_k.

For the general case a similar reasoning should be used to derive the formula for the *r:th* order inclusion probabilities, for $i_1 < ... < i_r$ and $n \geq r$,

$$\breve{\pi}_{i_1,...,i_r} = \frac{p_{i_1}, ..., p_{i_r} S_{n-r}^{N-r}([p_{i_1}, ..., p_{i_r}]^c)}{S_n^N(p_1, ..., p_N)}.$$

For larger populations this method can be slow and ineffective due to the amount of operations required to complete the recursion process. Therefore we present below an alternative method to calculate third order inclusion probabilities that is a straightforward generalisation of the second order methods in (1.3) and (1.4).

Let $\breve{\pi}_{ij\backslash k}$ denote the probability that elements *i,j* but not element k belong to the sample s, for a Conditional Poisson sample of size n, $n \geq 3$. We denote the odds by $\gamma_i = p_i/(1-p_i)$. From equation (1.1) it may be deduced that,

$$\breve{\pi}_{ij\backslash k} = \frac{\sum_{s \in A_n^{ij\backslash k}} \frac{p_j}{(1-p_j)} \prod_{h \in s'} p_h/(1-p_h)}{\sum_{s \in A_n} \prod_{l \in s} p_l/(1-p_l)},$$

where $A_n^{ij\backslash k}$ is the set of $s \in A_n$ with $ij \in s$, $k \notin s$, and s' is the set s excluding element j. Furthermore,

$$\breve{\pi}_{ij\backslash k} = \underbrace{\frac{\sum_{s \in A_n^{ik\backslash j}} \frac{p_k}{(1-p_k)} \prod_{h \in s*} p_h/(1-p_h)}{\sum_{s \in A_n} \prod_{l \in s} p_l/(1-p_l)}}_{\breve{\pi}_{ik\backslash j}} \frac{\gamma_j}{\gamma_k} = \breve{\pi}_{ik\backslash j} \cdot \frac{\gamma_j}{\gamma_k}, \tag{2.2}$$

$s*$ is the set s excluding element k. On the other hand,

$$\begin{cases} \breve{\pi}_{ij} = \breve{\pi}_{ij\backslash k} + \breve{\pi}_{ijk} \\ \\ \breve{\pi}_{jk} = \breve{\pi}_{jk\backslash i} + \breve{\pi}_{ijk}. \end{cases} \tag{2.3}$$

By substituting (2.2) in (2.3),

$$\begin{cases} \breve{\pi}_{ij} = \frac{\gamma_j}{\gamma_k} \breve{\pi}_{ik\backslash j} + \breve{\pi}_{ijk} \\ \\ \breve{\pi}_{jk} = \breve{\pi}_{ik\backslash j} + \breve{\pi}_{ijk}. \end{cases}$$

Combining the two equations, we get

$$\breve{\pi}_{ijk} = \frac{\gamma_j \breve{\pi}_{ik} - \gamma_k \breve{\pi}_{ij}}{\gamma_j - \gamma_k} \tag{2.4}$$

in case $\gamma_j \neq \gamma_k$.

For the case $\gamma_j = \gamma_k$ we need:

Lemma 1. *For a sampling design with fixed sized n and for $i \neq j$,*

$$\sum_{k:k\notin\{i,j\}} \pi_{ijk} = (n-2)\pi_{ij} \tag{2.5}$$

where π_{ijk} and π_{ij} are the obviously defined multiple inclusion probabilities.

Proof: The inclusion of a given element k in a sample s may be regarded as a random event indicated by the random variable I_k defined as

$$I_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{if not} \end{cases}$$

By the relations $\sum_U I_k = \sum_U \pi_k = n$ and $E[I_i I_j] = \pi_{ij}$, see Särndal *et al.* (1992), page 36-38, we then get

$$\begin{aligned} \sum_{k:k\notin\{i,j\}} \pi_{ijk} &= \sum E(I_i I_j I_k) = E[I_i I_j (\sum_U I_k - I_i - I_j)] \\ &= E[I_i I_j (n - I_i - I_j)] = nE[I_i I_j] - E[I_i^2 I_j] - E[I_j^2 I_i] \\ &= (n-2)\pi_{ij}. \end{aligned}$$

Now fix k_0 such that $\gamma_{k_0} = \gamma_j$. From the lemma we get that:

$$(n-2)\breve{\pi}_{ij} = \sum_{k:k\notin\{i,j\}} \breve{\pi}_{ijk} = (\sum_{k:k\notin\{i,j\},\gamma_k\neq\gamma_j} \breve{\pi}_{ijk}) + h\breve{\pi}_{ijk_0},$$

where h is the number of elements $k \notin \{i,j\}$ and $\gamma_{k_0} = \gamma_j$ (or equivalently $p_{k_0} = p_j$). Thus,

$$\breve{\pi}_{ijk_0} = \frac{(n-2)\breve{\pi}_{ij} - \sum_{k:k\notin\{i,j\},\gamma_k\neq\gamma_j} \breve{\pi}_{ijk}}{h}. \tag{2.6}$$

Note that the number of unordered triplets in a sample of size n is always $\binom{n}{3}$ and therefore $\sum_{i<j<k} \breve{\pi}_{ijk} = \binom{n}{3}$ can be used as a checksum.

In the general case an analogous reasoning can be done to show that:

Theorem 1. *For a Conditional Poisson sample of size n, the inclusion probabilities of r:th order, $n \geq r$ and $i_1 < ... < i_r$, are*

$$\breve{\pi}_{i_1,...,i_r} = \frac{\gamma_{i_{r-1}} \breve{\pi}_{i_1,...,i_{r-2},i_r} - \gamma_{i_r} \breve{\pi}_{i_1,...,i_{r-1}}}{\gamma_{i_{r-1}} - \gamma_{i_r}}$$

in case $\gamma_{i_{r-1}} \neq \gamma_{i_r}$. For the case $\gamma_{i_{r-1}} = \gamma_{i_r}$, fix i_0 such that $\gamma_{i_{r-1}} = \gamma_{i_0}$, then

$$\breve{\pi}_{i_1,...,i_{r-1},i_0} = \frac{(n-r)\breve{\pi}_{i_1,...,i_{r-1}} - \sum_{i_r:\gamma_{i_{r-1}} \neq \gamma_{i_r}, i_r \neq i_1,...,i_{r-1}} \breve{\pi}_{i_1,...,i_r}}{h}.$$

where h is the number of elements $i_r \neq i_1, ..., i_{r-1}$ such that $\gamma_{i_0} = \gamma_{i_{r-1}}$ or equivalently $p_{i_0} = p_{i_{r-1}}$. The control sum for the general case is

$$\sum_{i_1 < ... < i_r} \breve{\pi}_{i_1,...,i_r} = \binom{n}{r}.$$

Remark: In a computer implementation for the algorithm in the theorem, one shall permute the individuals and choose $\gamma_{i_{r-1}}$ and γ_{i_r} such that the difference $\gamma_{i_{r-1}} - \gamma_{i_r}$ become large enough to avoid values close to zero in the denominator.

Acknowledgements

I wish to thank Professor Olle Nerman for his valuable help throughout this work.

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**, pp. 457-469.

Dupačová, J. (1979). A note on Rejective Sampling, *Contributions to Statistics*, Jaroslav Hájek Memorial Volume, Reidel, Holland and Academia, Prague, 71-78.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling with Varying

Probabilities from a Finite Population, *Annals of Mathematical Statistics*, **35**, 1491-1523.

Hájek, J. (1981). *Sampling from a Finite Population*, Marcel Dekker, New York.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansions*, Springer Verlag, New York.

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

Thompson, M.E. (1997). *Theory of Sample Surveys*, Chapman and Hall, London. '

Paper D

Order Sampling Design with Prescribed Inclusions

Nibia Aires, Johan Jonasson, Olle Nerman

15 November 1999

Abstract

Order Sampling with Fixed Distribution Shape is a class of sampling schemes with inclusion probabilities approximately proportional to given size measures. Methods were provided to compute the exact first and second order inclusion probabilities numerically when the distribution shape is of the Pareto type. Procedures were also provided for this case to adjust the parameters to get predetermined inclusion probabilities. In this paper we prove the existence and uniqueness of a solution for the latter problem, in general for any order sampling of fixed distribution shape.

Keywords: Sampling Theory, Order Sampling, Pareto πps Sampling

AMS 1991 subject classification: 62D05, 65U05

1. Introduction

An asymptotic theory for order sampling is given by Rosén (1997a, 1997b). He introduces the notion of Order Sampling with Fixed Distribution Shape and he also shows that the Pareto πps sampling design (PPS), which is a scheme that belongs to this class, is asymptotically uniformly optimal among the schemes in this class which have inclusion probabilities asymptotically proportional to given size measures (πps). Algorithms to find numerical values of the exact first and second order inclusion probabilities were proposed in Aires (1999), for PPS, as well as two procedures to adjust the parameters when the exact inclusion probabilities are given. However, the latter were based on a conjecture assuming the existence of a unique solution given the exact inclusion probabilities. In this paper we prove this conjecture, i.e. that these parameters can be arbitrarily prescribed for any order sampling with fixed distribution shape. The argument will follow the same line as in Jonasson *et al.* (1997), where a corresponding theorem was proved for Conditional Poisson Sampling procedures. However since during the publication procedure an old quite different proof of that theorem was found, see Dupačová (1979), the proof was never published. Thus we shall here give a full account of all the quite involved arguments.

2. Preliminaries

2.1. Order Sampling

Consider a population $U = \{1, ..., N\}$. To each unit i in the population is associated a probability distribution $F_i(t)$ with density $f_i(t)$, $0 \leq t < \infty$. To realize an Order Sampling scheme of sample size n, with $n < N$, we consider independent ranking variables $Q_1, Q_2, ..., Q_N$ with continuous distribution functions $F_1, F_2, ..., F_N$. The units with the n smallest Q-values constitute the sample s. The idea of Order Sampling originates from the special case where all Q_i are uniform in which case the inventor Esbjörn Ohlsson (1995), named the resulting procedure Sequential Poisson Sampling. The idea was further developed by Rosén (1997a, 1997b), who also introduced the PPS scheme and stated the following definition of Order Sampling with fixed distribution shape:

Definition 1. *Let $H(t)$ be a probability distribution with density $h(t) > 0$, $t \in [0, c]$, where $c = \infty$ is allowed; and let $\theta = (\theta_1, ..., \theta_N)$, $\theta_i > 0$, be real numbers. An Order Sampling scheme with $F = (F_1, F_2, ..., F_N)$, is said to have fixed order distribution shape $H(t)$ and intensities θ, if either, and hence both, of the following two equivalent conditions are met.*

1. *The ranking variables $Q_1, ..., Q_N$ are of type $Q_i = V_i/\theta_i$, $i = 1, ..., N$, where $V_1, ..., V_N$ are iid random variables with common distribution H.*

2. *The order distributions are $F_i(t) = H(t \cdot \theta_i)$, $0 \leq t < \infty$, $i = 1, ..., N$.*

Thus an important subclass is derived by the assumption that all F_i belong to the same scale family and in case these distribution functions are defined as $F_i(t) = \theta_i t/(1 + \theta_i t)$, $\theta_i > 0$, we have a PPS procedure. PPS based Horvitz-Thompson estimators are asymptotically uniformly optimal among order sampling schemes with inclusion probabilities proportional to given size measures and with fixed distribution shape, cf. Rosén (1997b). To calculate the exact inclusion probabilities for an element $i \in U$ under order sampling design, we calculate first the distribution functions of order statistics of a sample of independent, but not necessarily identically, distributed random variables, see Aires (1999),

Lemma 1. *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 the recursive equation:*

$$F^N_n(t) = F^{N-1}_n(t) + F_N(t)[F^{N-1}_{n-1}(t) - F^{N-1}_n(t)], \tag{2.1}$$

where $F_0^N(t) = 1$, *for all* N *and* t.

Returning to the order sampling procedure, the probability of element N belonging to the sample s is

$$\begin{aligned} \pi_N &= P(i \in s) = P\left(Q_{(n)}^{N-1} > Q_N\right) \\ &= \int_0^\infty (1 - F_n^{N-1}(t)) f_N(t) dt. \end{aligned} \tag{2.2}$$

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. The recursion in (2.1) and the integral in (2.2) can be implemented to calculate π_i numerically with high precision.

Let g_i be π_i as defined in (2.1) and (2.2) seen as a function of the θ-parameters,

$$g_i(\theta_1, ..., \theta_N) = \int_0^\infty (1 - F_n^{N-1}(\{\theta_i\}^c, t)) f_{\theta_i}(t) dt.$$

This means that F_n^{N-1} is the distribution function of $Q_{(n)}^{N-1}$ as in (2.1), $f_{\theta_i}(t)$, $i = 1, ..., N$ is the density of Q_i and $\{\theta_i\}^c$ is the set $\{\theta_1, ..., \theta_N\} \setminus \{\theta_i\}$.

In Aires (1999), two methods are introduced to adjust the θ-parameters for the PPS design to solve the equation system:

$$\begin{cases} g_i(\theta_1, ..., \theta_N) = z_i, \\ \\ \sum_{i=1}^N F_i(1) = \sum_{i=1}^N H(\theta_i) = n \end{cases} \tag{2.3}$$

Here we show the existence and uniqueness of the solution of (2.3):

Theorem 1. *For any set of values of* $z_1, ..., z_N$ *such that* $0 \leq z_i \leq 1$ *and* $\sum_{i=1}^N z_i = n$, *the system of equations (2.3) has a unique solution for any fixed* $\theta_N \in [0, \infty)$. *Furthermore it is always possible to choose* θ_N *such that this solution satisfies* $\sum_{i=1}^N F_i(1) = \sum_{i=1}^N H(\theta_i) = n$.

3. Proof of the theorem

As already mentioned, the technique we use here is based on ideas in an earlier proof of a corresponding result for Conditional Poisson Sampling procedures Jonasson *et al.* (1997).

Remark: Assuming that $0 < z_i < 1$ for every i is no less of generality; extending to the situation where some z_i's are 0 or 1 is a trivial task.

For simplicity of notation, write (2.3) as

$$\begin{aligned} &g_1(\theta_1, ..., \theta_N) = z_1 \qquad (e1) \\ &\vdots \\ &g_N(\theta_1, ..., \theta_N) = z_N \qquad (eN) \end{aligned}$$

Note first that all g_i's are continuous and strictly increasing in θ_i. To see this, we observe from the recursive equation (2.1) in Lemma 1 that all $F_n^N(t)$ are continuous and strictly increasing in $[0, c]$.

Furthermore, from Definition 1,

$$P(i \in s) = P\left(Q_i < Q_{(n)}^{N-1}\right) = P\left(\frac{V_i}{\theta_i} < Q_{(n)}^{N-1}\right) = P\left(V_i < \theta_i Q_{(n)}^{N-1}\right)$$

is increasing in θ_i.

Now to solve the system of equations we will calibrate the θ_i's one by one and keep the remaining ones fixed. We assume without loss of generality that $z_1 \leq z_2 \leq ... \leq z_N$ and that $g_1, ..., g_N$ are ordered in the same order as $\theta_1, ..., \theta_N$. Also $\sum_{i=1}^N g_i(\theta_1, ..., \theta_N) = n$ is always automatically true.

First, we fix $(\theta_2, ..., \theta_N)$ at arbitrary values in $[0, \infty)$. Since $g_1(\theta_1, ..., \theta_N)$ is increasing in θ_1 and tends to 0 as $\theta_1 \to 0$ and to 1 as $\theta_1 \to \infty$ there is by continuity a unique θ_1 satisfying (e1). This θ_1 can be written as a function $g_1^1(\theta_2, ..., \theta_N)$, where g_1^1 is of course continuous. Inserting this into (e2) yields the equation

$$z_2 = g_2(g_1^1(\theta_2, ..., \theta_N), \theta_2, ..., \theta_N)$$

which we next, for arbitrary $\theta_3, ..., \theta_N$, will show has a solution

$$\theta_2 = g_2^2(\theta_3, ..., \theta_N).$$

Since g_1 is increasing in θ_1 and decreasing in θ_2 it follows that g_1^1 is increasing in θ_2.

Furthermore, $g_1(g_1^1(\theta_2, ..., \theta_N), \theta_2, ..., \theta_N) = z_1$ and $g_j(g_1^1(\theta_2, ..., \theta_N), \theta_2, ..., \theta_N)$ is decreasing in θ_2 for $j \geq 3$, as g_1^1 is increasing in θ_2. Thus it follows that

$$g_2 = n - z_1 - g_3 - ... - g_N$$

is increasing in θ_2 and by continuity of g_1^1 it is also continuous in θ_2. To see that z_2 is in the range of this function, we recall the assumption that $z_1 \leq z_2 \leq ... \leq z_N$. Letting $\theta_2 \to \infty$, which means that eventually $\theta_2 \geq \max(\theta_3, ..., \theta_N)$, we then have that $g_2 = \max(g_2, ..., g_N) \geq (n - z_1)/N \geq z_2$.

On the other hand letting $\theta_2 \to 0$, will either imply that $\theta_2 \leq g_1^1(\theta_2, ..., \theta_N)$ for some θ_2 so that $g_2 \leq g_1 = z_1 \leq z_2$ for this θ_2 or that g_1^1 will also tend to zero. If $n \leq N-2$ the latter would mean that $g_1 \to 0$, a contradiction, and if $n = N-1$ it would imply that $g_j \to 1$, $j \geq 3$, so that $g_1 + g_2 = z_1 + g_2 \to 1$. But since $n = N-1$ we have that $z_1 + z_2 > 1$, so that $g_2 < z_2$ for small enough θ_2 as desired. Thus, there is an unique pair $\theta_2 = g_2^2(\theta_3, ..., \theta_N)$ and

$$\theta_1 = g_1^2(\theta_3, ..., \theta_N) = g_1^1(g_2^2(\theta_3, ..., \theta_N), \theta_3, ..., \theta_N),$$

so that (e1) and (e2) are both satisfied. Moreover, observe that g_1^2 and g_2^2 are continuous.

Now, suppose that the procedure indicated above has been carried out for $l = 2, 3, ..., k$, $k < N-1$, so that at each step we have found unique values of $\theta_1, ..., \theta_l$ as continuous functions of $g_1^l, ..., g_l^l$ of $\theta_{l+1}, ..., \theta_N$ so that (e1),...,(el) are satisfied. It is then straightforward to carry out step $l = k+1$, if we verify first that $g_1^k, ..., g_k^k$ are increasing functions of θ_{k+1}. To see this, we assume that increasing θ_{k+1} to a larger value θ'_{k+1} causes g_i^k to decrease for $i \in B$, where B is a nonempty subset of $\{1, ..., k\}$, and the others to increase or stay fixed then the sum $\sum_{i \in B} g_i$ would decrease. To see that this cannot happen we proceed in two steps. First we increase θ_{k+1} and θ_i for $i \notin B$, $i \leq k$, to their larger new values and keep all the other θ_i's at the original values. Then all g_i's for $i \in B$ decrease. Second we change the θ_i's for $i \in B$ to their smaller new values. This causes all g_i's for $i \in \{1, ..., N\} \setminus B$ to increase and thus $\sum_{i \in B} g_i$ to decrease even further as $\sum_{i=1}^N g_i$ always remains equal to n.

The special treatment of the case $n = N-1$ above must be extended to the cases $n \geq N-k$; let θ_{k+1} tend to zero and assume that g_j^k also tends to zero for $j = 1, ..., k$, then, for such an n, g_i, $i = k+2, ..., N$ all tend to one, which implies that $g_{k+1} < z_{k+1}$ for small enough θ_{k+1} in the same way as for $l = 2$.

Thus we can proceed inductively to get, for each fixed θ_N, uniquely determined values of $\theta_1, ..., \theta_{N-1}$ satisfying (e1),...,(eN-1). Since $\sum_{i=1}^N g_i = n$, (eN) will be automatically satisfied. However, since all θ_i's, $i = 1, ..., N-1$ are increasing functions of θ_N, it is clear that $\sum_{i=1}^N g_i$ is increasing in θ_N. Letting $\theta_N \to 0$ implies that this sum also tends to 0 as $\theta_1 \leq ... \leq \theta_N$. On the other hand we can at this stage equally well regard $\theta_2, ..., \theta_N$ as functions of θ_1 and letting $\theta_1 \to \infty$ implies that $\sum_{i=1}^N g_i \to N$ by the same reason. It is easy to see that θ_N plays the role of an arbitrary scale parameter so that this sum equals n.

Now let $\theta_1, ..., \theta_{N-1}$ correspond to $\theta_N = 1$, then $(\theta_1 \cdot \kappa, ..., \theta_{N-1} \cdot \kappa, \kappa)$ will correspond to $\theta_N = \kappa$. And for $\theta_N = \kappa$, we get $\sum_{i=1}^N F_i(1) = \sum_{i=1}^N H(\theta_i \cdot \kappa)$, this sum is continuous and strictly increasing, from 0 to N, on the interval $\kappa \in [0, \infty]$, so there exist a unique κ such that this sum is n.

Remark: If $c < \infty$, we can not use $H(\theta_i)$ as unique parameters, since we can not discriminate between all $\theta_i \geq c$. However observe that if $c = \infty$ we may use $\lambda_i = H(\theta_i)$ as an alternative parametrisation. This parametrisation plays a fundamental role in an effective algorithm for finding the parameters in the Pareto case, Aires (1999). We conjecture that a direct generalisation should work well also in the general case.

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**, pp. 457-469.

Dupačová, J. (1979). A note on Rejective Sampling, *Contributions to Statistics*, Jaroslav Hájek Memorial Volume, Reidel, Holland and Academia, Prague, 71-78.

Jonasson, J. and Nerman, O. (1997). On Maximum Entropy πps-Sampling with Fixed Sample Size, Dept. of Mathematics, Göteborg University and Chalmers University of Technology, S-412 96 Göteborg, Sweden.

Ohlsson, E. (1995). Sequential Poisson Sampling, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Report No. 182.

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, 1997.

Paper E

On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling

Nibia Aires and Bengt Rosén

Chalmers University of Technology & Statistics Sweden

January 2000

1. Introduction and outline

A means for utilising auxiliary information in surveys is to sample with inclusion probabilities proportional to given size values, to use a πps design, preferably with fixed sample size. A candidate in that context is **Pareto πps**, introduced independently by Rosén (1997) and Saavedra (1995). This scheme has, as accounted for in Rosén (1997), many attractive properties, notably simple sample selection, good estimation accuracy, simple variance estimation and simple procedures for coordination of samples by permanent random numbers.

Pareto πps was derived by limit considerations, and works with some approximation. In particular, desired and factual inclusion probabilities do not agree exactly. Rosén (2000) proved, though, that they under very general conditions are asymptotically (as the sample size tends to infinity) equal. Numerical investigations by Rosén (2000) and Aires (1999, 2000) indicated that the convergence is rapid. These studies were too limited, though, to allow for general conclusions on how well desired inclusion probabilities are approximated by the factual ones. This paper reports on a much more extensive numerical study, in which the chief tool has been the algorithm in Aires (1999) for computation of Pareto πps inclusion probabilities.

The problem of how well desired inclusion probabilities are approximated has per se mainly basically theoretical interest. However, as is emphasised in the following, there is close connection between approximation accuracy for inclusion probabilities and estimator bias, the latter being an issue of great practical relevance. The convergence of inclusion probabilities implies that estimator bias is asymptotically negligible. Its magnitude for finite samples has been an open question, though. The chief aim in this paper is to enlighten this problem. The main conclusion is, somewhat sweepingly formulated, that the bias is negligible in practical survey situations.

The paper is organised as follows. Sections 2 and 3 are expository and review some basics on πps sampling in general respectively on Pareto πps.

Measures of approximation accuracy for inclusion probabilities and estimator bias are introduced in Section 4. Section 5 specifies certain size value patterns which play a distinguished role in the numerical study. The detailed numerical findings are presented in Appendices 1 and 2, containing tables and graphs respectively. Recommendations for practical use of Pareto πps are formulated in Section 6.

2. Generalities on πps sampling

We consider probability sampling without replacement with fixed sample size from a population $U = (1, 2, ..., N)$, on which a study variable $y = (y_1, y_2, ..., y_N)$ is defined. A frame which one-to-one corresponds with the population units is available. It is presumed that the frame contains unit-wise auxiliary information $s = (s_1, s_2, ..., s_N)$, $s_i > 0$, $i \in U$, interpreted as **size values** which typically are positively correlated with the study variable.

A sampling design is a **strict πps scheme** if its factual inclusion probabilities π_i, $i \in U$, are the following **desired inclusion probabilities** $\lambda = (\lambda_1, \lambda_2, ..., \lambda_N)$, n standing for sample size,

$$\lambda_i = n \cdot s_i / \sum_{j=1}^{N} s_j, \; i = 1, 2, ..., N. \tag{2.1}$$

Remark 1. *Formula (2.1) can lead to λ-values exceeding 1, which is incompatible with being probabilities. If so, the usual "adjustment" is to assign the units with largest size values to a "sample for certain" stratum. A πps sample is then drawn from the remaining units (with remaining sample size). In the sequel is presumed that* $0 < \lambda_i < 1$, $i \in U$.

As stated, a strict πps scheme is characterised by the relation $\pi_i = \lambda_i$, $i \in U$. We will be more generous, though, and accept a sampling scheme as a **πps scheme** if (2.2) below is met,

$$\pi_i \approx \lambda_i \tag{2.2}$$

holds with good approximation for $i = 1, 2, ..., N$. In the strict πps case, the Horvitz-Thompson (HT) estimator for a total $\tau(\mathbf{y}) = y_1 + y_2 + ... + y_N$ is as stated below. As is well known, this estimator is unbiased.

$$\hat{\tau}(\mathbf{y})_{HT} = \sum_{i \in Sample} \frac{y_i}{\lambda_i}. \tag{2.3}$$

We presume that the estimator in (2.3) is used also under the more generous πps notion based on (2.2). Then it may have some bias, though. The

estimator is re-stated in (2.4) where it is denoted $\hat{\tau}(\mathbf{y})$, which will be useful a bit later on. $I_1, I_2, ..., I_N$ denote the sample inclusion indicators, i.e. $I_i = 1$ if unit i is selected to the sample and $= 0$ otherwise.

$$\hat{\tau}(\mathbf{y}) = \sum_{i \in Sample} \frac{y_i}{\lambda_i} = \sum_{i=1}^{N} \frac{y_i}{\lambda_i} I_i. \tag{2.4}$$

3. On Pareto πps

3.1. Definition

Definition 1. *The Pareto πps scheme with size values $s = (s_1, s_2, ..., s_N)$ and* sample size *n generates a sample by the following steps.*

1. *The desired inclusion probabilities $\lambda = (\lambda_1, \lambda_2, ..., \lambda_N)$ are computed by (2.1).*

2. *Independent random variables $R_1, R_2, ..., R_N$ with uniform distribution on* $[0, 1]$ *are realized, and* ranking variables *Q are computed as follows,*

$$Q_i = \frac{R_i \cdot (1 - \lambda_i)}{\lambda_i \cdot (1 - R_i)}, \tag{3.1}$$

   $i = 1, 2, ..., N$.

3. *The sample consists of the units which have the n smallest Q-values.*

It is by no means obvious that the above scheme actually is a πps scheme (in the (2.2) sense). However, Rosén (2000) proved that (2.2) holds with asymptotic (as $n \to \infty$) equality.

As stated earlier, a main task for the present study was to find out how well approximation (2.2) works for finite Pareto πps samples. The central measure of approximation goodness will be the *maximal absolute relative bias* (for inclusion probabilities),

$$\Psi = \max \left| \frac{\pi_i}{\lambda_i} - 1 \right|, i = 1, 2, ..., N. \tag{3.2}$$

Ψ is a natural performance measure in the approximation problem, which is rather theoretical, though. However, Ψ also has considerable practical interest due to fact that there is close connection between Ψ and the magnitude of estimator bias, which is discussed next.

3.2. On estimator bias for Pareto πps

We presume that the sample is drawn by Pareto πps and that the estimator $\hat{\tau}(\mathbf{y})$ in (2.4) is used. As in Section 2, we confine the estimation considerations to the "fundamental" problem, estimation of a population total. Since (2.2) holds, $\hat{\tau}(\mathbf{y})$ is afflicted with some bias. To get an expression for it, we take expectation in (2.4),

$$E[\hat{\tau}(\mathbf{y})] = \sum_{i=1}^{N} \frac{y_i}{\lambda_i} \cdot E[I_i] = \sum_{i=1}^{N} y_i \cdot \frac{\pi_i}{\lambda_i} = \tau(\mathbf{y}) + \sum_{i=1}^{N} y_i \cdot \left(\frac{\pi_i}{\lambda_i} - 1\right). \quad (3.3)$$

Hence, the *bias for the estimator* $\hat{\tau}(\mathbf{y})$ is

$$E[\hat{\tau}(\mathbf{y})] - \tau(\mathbf{y}) = \sum_{i=1}^{N} y_i \cdot \left(\frac{\pi_i}{\lambda_i} - 1\right). \quad (3.4)$$

Formulas (3.3) and (3.4) yield, the *absolute relative bias for the estimator* $\hat{\tau}(\mathbf{y})$ is

$$\left|\frac{E[\hat{\tau}(\mathbf{y})] - \tau(\mathbf{y})}{\tau(\mathbf{y})}\right| = \left|\sum_{i=1}^{N} y_i \cdot \left(\frac{\pi_i}{\lambda_i} - 1\right) / \tau(\mathbf{y})\right| \leq \Psi \cdot \left(\sum_{i=1}^{N} |y_i| / \tau(\mathbf{y})\right). \quad (3.5)$$

If the study variable $\mathbf{y}$ takes only non-negative values, as is the case in most practical surveys, the last factor in (3.5) equals 1. Hence, for a non-negative study variable $\mathbf{y}$, the absolute relative bias for $\tau(\mathbf{y})$ is

$$\left|\frac{E[\hat{\tau}(\mathbf{y})] - \tau(\mathbf{y})}{\tau(\mathbf{y})}\right| \leq \Psi. \quad (3.6)$$

Remark 2. *The bounds in (3.5) and (3.6) are often conservative for the following reasons. (i) They disregard cancellation effects due to alternating signs of* $\pi_i/\lambda_i - 1$. *(ii) All discrepancies* $|\pi_i/\lambda_i - 1|$ *do not have the maximal value* Ψ. *The bounds can be attained, though, e.g. with* $y_i = 0$ *for* i *with* $|\pi_i/\lambda_i - 1| < \Psi$, *and* $y_i = sign(\pi_i/\lambda_i - 1)$ *for* i *with* $|\pi_i/\lambda_i - 1| = \Psi$.

3.3. Chief questions in the numerical study

A pair $(N;\mathbf{s})$ of a population size N with size values $\mathbf{s} = (s_1, s_2, ..., s_N)$ is referred to as a **size measure situation**. A Pareto πps scheme is specified by $(N;\mathbf{s})$ and the sample size n. When we want to emphasise dependence on

one or more of these parameters, we use notation as Pareto $\pi ps(N, \mathbf{s}, n)$ or Pareto $\pi ps(\mathbf{s})$, $\lambda_i(n)$, $\lambda_i(n;\mathbf{s})$, $\lambda_i(n;N;\mathbf{s})$ and $\pi_i(n)$, $\pi_i(n;\mathbf{s})$, $\pi_i(n;N;\mathbf{s})$ for desired and factual inclusion probabilities. Analogously, Ψ in (3.2) is often elaborated to

$$\Psi(n;N;\mathbf{s}) = \max_i \left| \frac{\pi_i(n;N;\mathbf{s})}{\lambda_i(n;N;\mathbf{s})} - 1 \right|. \tag{3.7}$$

The chief problems that are addressed are stated in Question 1 and Question 2 below. Note that Question 1 is a "converse" to Question 2. For a specific size value situation $(N;\mathbf{s})$,

Question 1. *How large, at most, is $\Psi(n;N;\mathbf{s})$ for a specific sample size n?*

Question 2. *Which sample sizes n imply $\Psi(n;N;\mathbf{s}) \leq \beta$ for a specified $\beta > 0$?*

In the first round Questions 1 and 2 relate to the approximation (2.2), how well the factual inclusion probabilities approximate the desired ones. However, by virtue of (3.5) and (3.6) answers to Questions 1 and 2 also provides information on relative estimation bias. In the sequel **the approximation problem** refers to both these aspects, estimator bias as well as discrepancy between factual and desired inclusion probabilities.

To the best of our understanding it is in vain to hope for answers to Questions 1 and 2 via analytical formulas. One has to be content with (fair) coarse answers derived by numerically demanding computation efforts. The employed numerical algorithm is described next.

3.4. The computation algorithm

The chief work in deriving answers to Questions 1 and 2 consisted of computation of $\pi_i(n;N;\mathbf{s})$, $i = 1, 2, ..., N$, for a rich set of values for $(N, \mathbf{s})$ and n. For that the core tool was the algorithm for computation of Pareto πps inclusion probabilities which is derived and justified in Aires (1999). To give an idea of the numerical efforts, a sketch of the algorithm is presented below. It describes computation of $\pi_i(n;N;\mathbf{s})$ for $i = N$. π_i-values for a general i were computed by appropriate re-labelling of the population units.

Computation algorithm for Pareto inclusion probabilities

The given quantities are N, $\mathbf{s}$ and n.
Step 1: Compute $\lambda_i(n, \mathbf{s})$ by (2.1).

Step 2: For a mesh M of t-values, which is fine enough to yield desired precision in the numerical integration in Step 3, determine $\{F_n^N(t),\ t \in M\}$ by the double recursion (in n and N),

$$F_n^N(t) = F_n^{N-1}(t) + F_N(t)[F_{n-1}^{N-1}(t) - F_n^{N-1}(t)], \tag{3.8}$$

where $F_0^N(t) = 1$, for all N *and* $0 \leq t < \infty$. $F_N(t) = \theta_N \cdot t/(1 + \theta_N \cdot t)$ with $\theta_N = \lambda_N/(1 - \lambda_N)$. F_N is the distribution function of Q_N and it is not hard to see that $F_n^N(t)$, $0 \leq t < \infty$, is the distribution function of the n:th order statistic of $Q_{(1)}, ..., Q_{(N)}$. $1/(1+\theta_N \cdot t)^2$ is the density of Q_N and (3.9) follows by a straight forward conditioning argument.

Step 3 : Compute, by numerical integration

$$\pi_N = \theta_N \cdot \int_0^\infty (1 - F_n^{N-1}(t))/(1 + \theta_N \cdot t)^2 dt. \tag{3.9}$$

4. Bounds employed in the approximation problem

4.1. Some definitions

In the sequel size measures $\mathbf{s} = (s_1, s_2, ..., s_N)$ are presumed to be normed so that average size is 1, i.e. so that (4.1) below holds,

$$\frac{1}{N} \cdot \sum_{i=1}^{N} s_i = 1. \tag{4.1}$$

A normed $\mathbf{s}$ is called a **size value pattern**. Set

$$s_{min} = \min\{s_i : i = 1, 2, ..., N\},\ s_{max} = \max\{s_i : i = 1, 2, ..., N\}. \tag{4.2}$$

As stated in Remark 1, it is presumed that all λ_i given by (2.1) are smaller than 1. This lays the following constraint on sample sizes n, where $[\cdot]$ denotes integral part and - "less than",

$$n \leq n_m = n_m(N; \mathbf{s}) = [N/s_{max}-]. \tag{4.3}$$

The quantity n_m in (4.3) is called the **maximal sample size** in situation $(N; \mathbf{s})$. An n which satisfies (4.3) is said to be an **admissible sample size** in situation $(N; \mathbf{s})$. λ-values close to 1 may lead to "capricious" samples, which can be avoided by prescribing that $\lambda_i \leq \alpha,\ i = 1, 2, ..., N$ for some specified $\alpha < 1$. $\alpha = 0.9,\ 0.8$ were considered in the numerical context. The **α-maximal sample size** $n_{m,\alpha}$ and **α-admissible sample sizes** are determined by,

$$n \leq n_m = n_{m,\alpha}(N; \mathbf{s}) = [\alpha \cdot N/s_{max}]. \tag{4.4}$$

In Section 6 $n_{m,\alpha}$ is also used as a means for stating conditions to the effect that a sample size must not be "too large". For that purpose also $\alpha = 0.5$ was considered. We shall relate approximation error bounds to **size pattern families** of the following type

$$S(N;\gamma,\delta) = \{\mathbf{s} : \textit{the population size} = N, s_{min} = \gamma \textit{ and } s_{max} = \delta\},$$
$$0 < \gamma \leq 1 \leq \delta < \infty. \quad (4.5)$$

In words: a size pattern is in $\mathbf{S}(N;\gamma,\delta)$ if at least one population unit has normed size γ, at least one size δ and the others have size values in $[\gamma,\delta]$. This type of family is of interest for at least the following reasons.

1. When all size values are equal, $\delta = \gamma = 1$, Pareto $\pi ps(\mathbf{s})$ is nothing but simple random sampling, with $\pi_i(n) = \lambda_i(n) = n/N$, and the approximation (2.2) is perfect. Thus, for an approximation problem to be at hand, different size values must occur. $\mathbf{S}(N;\gamma,\delta)$ lays constraints on how different they may be. The smaller γ and the larger δ are, the more different are the size values.

2. It is simple to determine to which family $\mathbf{S}(N;\gamma,\delta)$ a size pattern belongs by computing the smallest and largest normed size values.

Since $n_m(N;\mathbf{s})$ and $n_{m,\alpha}(N;\mathbf{s})$ in (4.3) and (4.4) depend only on N and s_{max}, they are the same for all patterns in $\mathbf{S}(N;\gamma,\delta)$. We therefore use the following simpler notation; for $\mathbf{s} \in \mathbf{S}(N;\gamma,\delta)$ we write

$$n_m(N;\delta) \textit{ and } n_{m,\alpha}(N;\delta) \quad (4.6)$$

for $n_m(N;\mathbf{s})$ and $n_{m,\alpha}(N;\mathbf{s})$.

4.2. Bounding sequences

A size value situation $(N;\mathbf{s})$ determines a sequence $\{\Psi(n;N;\mathbf{s}) : n = 1,2,...,n_m\}$, with Ψ given by (3.7), which we refer to as the associated **Ψ-sequence**. Such sequences will play a central role in the subsequent considerations.

At the outset of this study we had various conjectures about the behaviour of Ψ-sequences. Many of these turned out to be wrong when confronted with numerical data. One was that all Ψ-sequences have bathtub shape. That this is not true in general is illustrated in Figure 4.1, which shows Ψ-sequences for three different size value patterns, all with $N = 100, \gamma = 0.5$ and $\delta = 2$. Their names, "boundary", "middle" and "even" are explained later on. A multitude of other Ψ-sequence graphs are presented in Appendix 2.

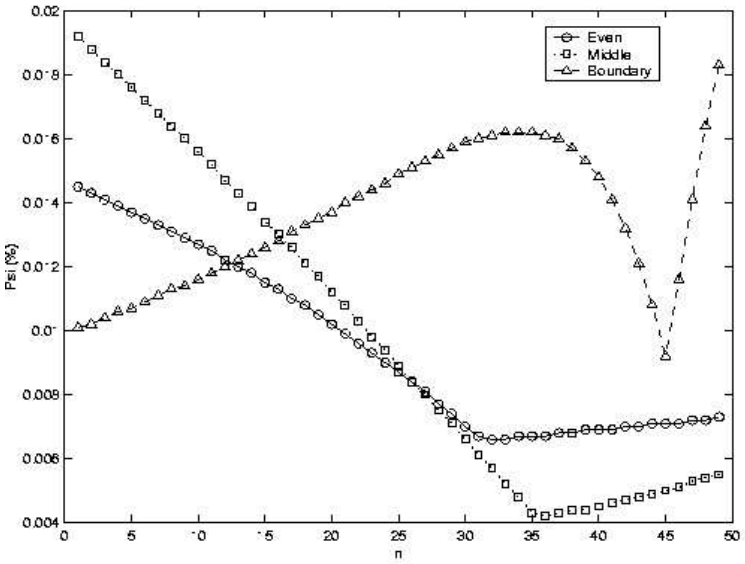

Figure 1: Ψ sequences for three size patterns

Our aim is to answer Questions 1 and 2 in terms of the parameters N, γ and δ. Figure 1 shows that there is no simple domination rule for Ψ-sequences for different size value patterns in the same family $\mathbf{S}(N;\gamma,\delta)$. As functions of n they can take turn to lie above each other. To find bounds which hold uniformly for N, γ and δ we must introduce envelope notions.

4.2.1. Ψ-envelope sequences

The Ψ-envelope sequence for the family $\mathbf{S}(N;\gamma,\delta)$, denoted $\Psi^*(\cdot\,;N;\gamma,\delta)$, is (recall (4.6)),

$$\Psi^*(n;N;\gamma,\delta) = \sup_{s\in S(N;\gamma,\delta)} \Psi(n;N;\mathbf{s}), n = 1,2,...,n_m(N;\delta). \qquad (4.7)$$

In words, $\Psi^*(\cdot\,;N;\gamma,\delta)$ means the maximal relative approximation error in (2.2) for a Pareto πps sample of size n selected from a population of size N with maximal and minimal normalised size values γ and δ. Technically

formulated, for $\mathbf{s} \in \mathbf{S}(N;\gamma,\delta)$ and $n \leq n_m(N;\delta)$:

$$\left|\frac{\pi_i(n,\mathbf{s})}{\lambda_i(n,\mathbf{s})} - 1\right| \leq \Psi^*(n;N,\gamma,\delta),\ i = 1,2,...,N. \tag{4.8}$$

Hence, knowledge of $\Psi^*(\cdot\,;N;\gamma,\delta)$ enables answers to Question 1. In fact, Ψ^* is the smallest possible upper bound sequence that works for all $\mathbf{s}$ in $\mathbf{S}(N;\gamma,\delta)$.

4.2.2. Quasi envelopes

The envelope (4.7) is defined in terms of suprema over the infinite family $\mathbf{S}(N;\gamma,\delta)$. To compute it in practice, one needs to know of a finite extremal sub-family of $\mathbf{S}(N;\gamma,\delta)$, by which we mean a finite sub-family with the same envelope. Regrettably, we cannot exhibit such a sub-family with mathematical rigour. However, we strongly believe, supported by numerical findings, that the following intuitive arguments lead to a "close to extremal" sub-family.

Numerically extremal size value patterns (as regards Ψ-values) are found among geometrically extremal patterns. In the latter category, the following three types of size value patterns come into mind.

1. Patterns with size values (fairly) **evenly spread** over $[\gamma,\delta]$.
2. Patterns with the majority of size values in the **middle** of $[\gamma,\delta]$,
3. Patterns with the majority of size values at the **boundaries** of $[\gamma,\delta]$,

Precise specifications of such patterns are given in Section 5, where they are denoted $\mathbf{s}(N;\gamma,\delta,\mathbf{e})$, $\mathbf{s}(N;\gamma,\delta,\mathbf{m})$ and $\mathbf{s}(N;\gamma,\delta,\mathbf{b})$, $\mathbf{e}$ for "even spread", $\mathbf{m}$ for "middle" and $\mathbf{b}$ standing for "boundary".

The quasi Ψ-envelope sequence for $\mathbf{S}(N;\gamma,\delta)$, denoted $\Psi^{**}(\cdot;N;\gamma,\delta)$ is

$$\begin{aligned}\Psi^{**}(n;N,\gamma,\delta) = \max\{&\Psi(n;\mathbf{s}(N,\gamma,\delta,\mathbf{e}),\Psi(n;\mathbf{s}(N,\gamma,\delta,\mathbf{m})),\\ &\Psi(n;\mathbf{s}(N,\gamma,\delta,\mathbf{b}))\},\ n = 1,2,...,n_m(N,\delta).\end{aligned} \tag{4.9}$$

Believing that $s(\mathbf{e}), s(\mathbf{m}), s(\mathbf{b})$ is a "close to extremal" sub-family of $\mathbf{S}(N;\gamma,\delta)$ we work under the following presumption in the sequel,

Conjecture 1. *$\Psi^{**}(\cdot\,;N,\gamma,\delta)$ yields good approximation of the true envelope $\Psi^*(\cdot\,;N,\gamma,\delta)$.*

Since Ψ^{**} is determined by just three size value patterns it is computable provided that a computation algorithm for $\Psi(n,N,\mathbf{s})$ for a given $\mathbf{s}$ and n is available, which it is by Section 3.4. Strictly mathematically, though, Conjecture 1 is only a conjecture, based on intuition and numerical support. We

made attempts to justify Conjecture 1 more rigorous by employing numerical optimisation programs to find extremal size value patterns. However, this approach turned out to be unfeasible, at least with the optimisation programs we tried.

4.2.3. Upper sequences

The quasi envelope may be quite irregular, wiggling up and down, as seen in Figure 1. To enable simple answers to Questions 1 and 2 we introduce coarser upper bound sequences (than the quasi envelope), called upper sequences, which are non-increasing functions of sample size.

The upper sequence $\Gamma(\cdot)$ for the family $\mathbf{S}(N;\gamma,\delta)$ is

$$\Gamma(n_0;N;\gamma;\delta) = \max\{\Psi^{**}(n;N;\gamma;\delta) : n_0 \le n \le n_m(N;\delta),\} \\ n_0 = 1,2,...,n_m(N,\delta). \quad (4.10)$$

In words, $\Gamma(n_0;N;\gamma;\delta)$ bounds the relative approximation error in (2.2) for a Pareto πps sample from a population of size N, with maximal and minimal normalised size values γ and δ, for all sample sizes $\ge n_0$. Under Conjecture 1 the following bound holds with good approximation. For $\mathbf{s} \in \mathbf{S}(N;\gamma,\delta)$,

$$\Psi(n;N,\mathbf{s}) \le \Gamma(n_0;N;\gamma;\delta),\ n_0 \le n \le n_m(N,\delta), n_0 = 1,2,...,n_m(N;\delta). \quad (4.11)$$

4.2.4. Upper sequence for α-admissible sample sizes

Since Pareto πps is based on limit considerations, one believes in the first round that conditions for good approximation basically should be of the type "provided the sample size is at least...". However, as seen from the graphs in Appendix 2, $\Psi^{**}(\cdot)$ often attains its largest values for large (admissible) sample sizes. As a consequence, sharp conditions for good approximation must also contain an ingredient of the type "provided the sample size is at most ...". This aspect will technically be handled as follows. An α is specified, $0 < \alpha < 1$, and used in conditions saying that sample sizes must not exceed $n_{m,\alpha}$ in (4.4), which is a way of saying "provided the sample size is at most ...". In line with this, the notion of upper sequence is extended as follows.

The upper sequence for α-admissible sample sizes is

$$\Gamma_\alpha(n_0;N;\gamma;\delta) = \max\{\Psi^{**}(n;N;\gamma;\delta) : n_0 \le n \le n_{m,\alpha}(N;\delta)\}, \\ n_0 = 1,2,...,n_{m,\alpha}(N;\delta). \quad (4.12)$$

Under Conjecture 1 also the following bounds hold with good approximation. For $\mathbf{s} \in \mathbf{S}(N;\gamma;\delta)$,

$$\Psi(n;N,\mathbf{s}) \leq \Gamma_\alpha(n_0;N;\gamma;\delta),\ n_0 \leq n \leq n_{m,\alpha}(N;\delta),$$
$$n_0 = 1,2,...,n_{m,\alpha}(N;\delta). \tag{4.13}$$

Remark 3. *The maximum operations in (4.10) and (4.12) add to what is said in Remark 2. In (4.11) and (4.13)* $\Gamma(n;N;\gamma;\delta)$ *and* $\Gamma_\alpha(n;N;\gamma;\delta)$ *may be quite conservative bounds for many sample sizes* n.

Appendix 1 presents numerical values for general upper sequences according to (4.10) as well as for α-admissible sample sizes, with $\alpha = 0.9, 0.8, 0.5$.

4.2.5. Sufficient sample sizes

Let β, $0 < \beta < 1$, be a specified tolerance level for the relative approximation error in (2.2). By disregarding the (mildly) approximative nature of the statement in Conjecture 1, answer to Question 2 is given by the smallest inverse n_0 such that $\Gamma(n_0;N;\gamma;\delta) \leq \beta$ which is called β-sufficient sample size for $\mathbf{S}(N;\gamma,\delta)$ and denoted $n_0(\beta)$. It informs about sample sizes which are large enough to guarantee approximation accuracy β. Formally

$$n_0(\beta) = \min\{n : \Gamma(n;N;\gamma;\delta) \leq \beta.\} \tag{4.14}$$

As discussed in Section 4.2.4, large sample sizes rather than small ones that jeopardise approximation accuracy, which is addressed by the following notion. The (β,α)-sufficient sample size for $\mathbf{S}(N;\gamma,\delta)$ denoted $n(\beta,\alpha)$, is the smallest sample size which guarantees that $\Psi^{**}(n;N;\gamma;\delta) \leq \beta$ for $n(\beta,\alpha) \leq n \leq n_{m,\alpha}(N;\delta)$. Formally,

$$n_0(\beta,\alpha) = \min\{n : \Gamma_\alpha(n;N;\gamma;\delta) \leq \beta\}. \tag{4.15}$$

The set of n-values over which minimum is taken in (4.14) and (4.15) may be empty. Then n_0 is set to **none**. Numerical β- and (β,α)-sufficient sample sizes are presented in Appendix 1.

4.2.6. Approximation accuracy and population size

A conjecture about Ψ-sequence behaviour which was supported by the numerical findings is formulated in Conjecture 2. Somewhat sweepingly expressed it says that approximation accuracy improves as population size increases. Still, also Conjecture 2 is a conjecture without a strict mathematical proof. It can be tested numerically in Appendix 1, though.

Conjecture 2. *For fixed* n, γ *and* δ *the values of upper sequences* $\Gamma(n;N;\gamma;\delta)$ *and* $\Gamma_\alpha(n;N;\gamma;\delta)$ *decrease as the population size* N *increases.*

4.2.7. Weak quasi envelopes

The quasi envelope in (4.9) dominates Ψ-sequences for all size pattern shapes. As illustrated by the graphs in Appendix 2, the "worst possible" in the boundary pattern unless when sample size is very small. Its Ψ-values mostly lie way above those for even spread and middle when sample size is "non-small". However, patterns of boundary type are unusual in practice where, at least we think so, most pattern shapes resemble even spread. Patterns in the last category will be said to lie **in the vicinity of even spread**. With this background we introduce the weak quasi envelope $w\Psi^{**}$, which takes into account only the even spread and middle patterns,

$$w\Psi^{**}(n) = \max\{\Psi(n; \mathbf{s}(N, \gamma, \delta, \mathbf{e})), \Psi(n; \mathbf{s}(N, \gamma, \delta, \mathbf{m}))\}, \\ n = 1, 2, ..., n_m(N, \delta). \qquad (4.16)$$

Moreover, weak upper sequences, weakly β-sufficient sample sizes and weakly (β, α)-sufficient sample sizes are defined in analogy with (4.10), (4.12) and (4.14)-(4.15) by using the weak quasi envelope $w\Psi^{**}$ instead of Ψ^{**}. Numerical values are presented in Appendix 1.

Remark 4. *In Section 4.2.4 is pointed at the circumstance that the (full) quasi envelope usually attains its largest values for large sample sizes. This growth depends mainly on contribution from size patterns of boundary type. The weak quasi envelope, which is not influenced by boundary type patterns, behaves as "expected", it decreases as sample size increases.*

5. The special size patterns

Here we give precise specifications of the size value patterns $\mathbf{s}$ in $\mathbf{S}(N; \gamma, \delta)$ which are mentioned in Section 4.2.2, boundary, middle and even patterns. Recall that (4.1) is presumed to hold, and the relation

$$\gamma = s_{min} \leq s_i \leq s_{max} = \delta, \; i = 1, 2, ..., N. \qquad (5.1)$$

Set

$$M_\gamma = N \cdot (\gamma - 1)/(\delta - \gamma), \; M_\delta = N \cdot (1 - \gamma)/(\delta - \gamma). \qquad (5.2)$$

Note the following relations

$$M_\gamma + M_\delta = N, \; \gamma \cdot M_\gamma + \delta \cdot M_\delta = N. \qquad (5.3)$$

M_γ and M_δ split into integral parts, N_γ and N_δ, and fractional parts, F_γ and F_δ as follows,

$$M_\gamma = [M_\gamma] + F_\gamma = N_\gamma + F_\gamma, \; M_\delta = [M_\delta] + F_\delta = N_\delta + F_\delta. \qquad (5.4)$$

Case 1 is said to be at hand if M_γ and M_δ both are integers. Then (5.3) takes the form,

$$N_\gamma + N_\delta = N,\ \gamma \cdot N_\gamma + \delta \cdot N_\delta = N. \tag{5.5}$$

Case 2 is said to be at hand if M_γ and M_δ not both are integers. Then none of them is an integer, since N_γ and N_δ add to an integer. Moreover, as is readily checked

$$F_\gamma + F_\delta = 1,\ N_\gamma + N_\delta = N - 1,\ N - (\gamma \cdot N_\gamma + \delta \cdot N_\delta) = \gamma \cdot F_\gamma + \delta \cdot F_\delta. \tag{5.6}$$

The special patterns are first specified, and then is shown that they satisfy (4.1) and (5.1).

The boundary pattern $\mathbf{s}(N; \gamma, \delta, \mathbf{b})$

It is presumed that N, γ and δ are such that $N_\gamma \geq 1$ and $N_\delta \geq 1$. For this size pattern N_γ units are given the s-value γ, and N_δ units the value δ. In Case 1 all units thereby get s-values. In Case 2 one unit remains, which is given the s-value

$$s_N = N - (\gamma \cdot N_\gamma + \delta \cdot N_\delta), \tag{5.7}$$

which by (5.2) also means

$$s_N = \gamma \cdot F_\gamma + \delta \cdot F_\delta.$$

The middle pattern $\mathbf{s}(N; \gamma, \delta, \mathbf{m})$

It is presumed that N, γ and δ are such that $\gamma \leq 1 - (\gamma + \delta - 2)/(N - 2) \leq \delta$. For this size pattern s-values are assigned as follows. The size values γ and δ are given to one unit each. All remaining units are given the s-value

$$s = 1 - (\gamma + \delta - 2)/(N - 2). \tag{5.8}$$

The even spread pattern $\mathbf{s}(N; \gamma, \delta, \mathbf{e})$

It is presumed that N, γ and δ are such that $N_\gamma \geq 2$ and $N_\delta \geq 2$. The s-values are allocated as follows.

$$If \qquad (1-\gamma)\cdot N_\gamma < (\delta-1)\cdot N_\delta \tag{5.9}$$

$$s_i = \gamma + (i-1)\cdot(1-\gamma)/(N_\gamma - 1), \tag{5.10}$$
$$i = 1, 2, ..., N_\gamma,$$
$$s_i = \delta - (N_\delta + N_\gamma - i)\cdot(1-\gamma)\cdot N_\gamma/[(N_\delta\cdot(N_\delta-1)], \tag{5.11}$$
$$i = N_\gamma + 1, N_\gamma + 2, ..., N_\gamma + N_\delta.$$

$$If \qquad (1-\gamma)\cdot N_\gamma \geq (\delta-1)\cdot N_\delta \tag{5.12}$$

$$s_i = \gamma + (i-1)\cdot(\delta-1)\cdot N_\delta/[N_\gamma\cdot(N_\gamma-1)], \tag{5.13}$$
$$i = 1, 2, ..., N_\gamma,$$
$$s_i = \delta - (N_\delta + N_\gamma - i)\cdot(\delta-1)/(N_\gamma-1), \tag{5.14}$$
$$i = N_\gamma + 1, N_\gamma + 2, ..., N_\gamma + N_\delta.$$

In Case 1 all units get s-values by (5.9)-(5.11). In Case 2 one unit remains, which is given the s-value in (5.7).

Thereby the size patterns are defined and the remaining task is to show that they satisfy (4.1) and (5.1). For the middle pattern the verifications are straightforward. The same holds for the boundary pattern, when noted that s_N in (5.7) is a linear combination of γ and δ and, hence, lies in the interval $[\gamma, \delta]$. For the even pattern we start with the case (5.9). Formulas (5.10) and (5.11) readily yield

$$\sum_{i=1}^{N_\gamma} s_i = \gamma\cdot N_\gamma + N_\gamma\cdot(1-\gamma)/2 \tag{5.15}$$

and

$$\sum_{i=1}^{N_\gamma+N_\delta} s_i = \delta\cdot N_\delta - N_\gamma\cdot(1-\gamma)/2, \tag{5.16}$$

which in turn yields

$$\sum_{i=1}^{N_\gamma+N_\delta} s_i = N_\gamma\cdot\gamma + \delta\cdot N_\delta. \tag{5.17}$$

In Case 1 the relation (5.1) follows from (5.17) and (5.5). In Case 2 it follows from the definition (5.7) of s_N. It remains to show (5.1). The s-values in (5.11) are generated as

$$s_i = \delta - (N_\delta + N_\gamma - i) \cdot (\delta - \epsilon)/(N_\delta - 1), \\ i = N_\gamma + 1, N_\gamma + 2, ..., N_\gamma + N_\delta, \tag{5.18}$$

where ϵ solves the equality version $(1 - \gamma) \cdot N_\gamma = (\delta - \epsilon) \cdot +N_\delta$, of the inequality in (5.9)

$$\epsilon = \delta - (1 - \gamma) \cdot N_\gamma / N_\delta. \tag{5.19}$$

(5.9) yields that $1 \leq \epsilon$, and (5.19) that $\epsilon \leq \delta$. Hence, all s-values in (5.11) lie in $[\gamma, \delta]$. The same holds for the s-values in (5.10). That s_N in (5.7) satisfies (5.1) follows from (5.7). Then case (5.12) can be treated quite analogously, and is left to the reader. This concludes the proof.

6. On the magnitude of estimator bias

6.1. Factors that affect the bias

6.1.1. Introduction

For a practitioner who considers to use Pareto πps, a crucial question is,

Question 3. *Will the estimator bias be negligible in my particular sampling-estimation situation?*

Answers to Question 3 in terms of practically available parameters are given in next section. They are with necessity a bit complex, since the approximation accuracy depends on several factors, notably those listed below and commented on thereafter.

- The study variable.
- The tolerance limit for "negligibility".
- The size value pattern.
- The population size.
- The sample size.

As regards the study variable, we confine to the case with non-negative variable, which is the typical situation in practice. Hence, subsequent statements about bias shall be interpreted according to (3.6). It is left to the reader to make appropriate modifications for situations with sign changing study variable.

6.1.2. Tolerance level for negligibility

There is of course no single answer to how large at most a "negligible" bias may be. It depends on the intended use of the statistic and the magnitude of other survey errors, notably the sampling error. We believe that most statisticians say that 1%, and even 2%, relative bias is negligible, a reason being that the sampling error commonly is a good deal larger.

6.1.3. Dependence on the size value variation

As said in Section 4.1, for an approximation problem to be at hand the size values must exhibit variation, having the aspects **range** and **shape**. The range is the interval $[s_{min}, s_{max}] = [\gamma, \delta]$. For a fixed size pattern shape $\Psi - values$ increase with range. Based on experience we believe that in practical surveys $s_{max} = \delta$ seldom is larger than 5 and $s_{min} = \gamma$ is seldom smaller than 0.1. Some motivation is given below.

The surveyor disposes of the size values, in the sense that "preliminary" values may be modified. If the frame comprises units with very small (preliminary) size values, such units are often either definition-wise excluded from the target population or given larger size values.

If size values vary very much over the entire population, there are often grounds for stratification by size before sampling, followed by selection of independent samples from the strata. (A typical example is given by an enterprise survey with number of employees as size. Then it is often natural to divide into strata of types "very big", "big", "medium" and "small" enterprises. Mostly the "very big stratum" is totally inspected). The strata then take population roles, and s_{max} and s_{min} in strata are considerably smaller/larger than in the entire population.

As regards size pattern shape our experience says, as stated in Section 4.2.7, that the boundary pattern, which is most adverse for good approximation, is very unusual in practice.

Table 1 below introduces, for later use, a broad categorisation of size value patterns.

Table 1: Some categories of size value patterns

	Category A	Category B	Category C	Category D
Size Pattern shape	In the vicinity of even spread	In the vicinity of even spread	No restriction.	No restriction.
$\mathbf{s_{max}}$	≤ 5	≤ 10	≤ 5	≤ 10
$\mathbf{s_{min}}$	≥ 0.1	≥ 0.05	≥ 0.1	≥ 0.05

Comment:

Category A. Most practical sampling situations are believed to fall in this category.

Category B. "Normal" pattern shape, while $\mathbf{s_{max}}$ and/or $\mathbf{s_{min}}$ may be extreme.

Category C. "Normal" $\mathbf{s_{max}}$ and $\mathbf{s_{min}}$, while pattern shape may be extreme (e.g. of boundary type).

Category D. Pattern shape as well as $\mathbf{s_{max}}$ and/or $\mathbf{s_{min}}$ may be extreme.

6.1.4. Dependence on population and sample sizes

As discussed in Section 4.2.6 approximation accuracy improves as population size increases (while γ, δ, α and n are fixed). Some of the figures in Table 2 below are based on extrapolation from available numerical data by employing the mentioned principle.

As regards dependence on sample size we refer to Sections 4.2.4 and 4.2.5.

6.2. Conditions for negligible estimator bias

The full numerical material to provide answers to Questions 1 and 2 is presented in the Appendices 1 and 2. It is somewhat difficult, though, to overview it as it stands in the appendices. The following Table 2 summarises the numerical findings at the prize of some coarsening. In some cases it may be helpful to consult the more detailed material in the appendices. Population sizes smaller than 25 were not considered in the study.

From Remarks 2 and 3 follows that the sufficient sample sizes in Table 2 are more or less conservative and, hence, "overly safety". In particular, one should not conclude that the bias necessarily is larger than "guaranteed" for sample sizes that are smaller than stated n_0-values.

Our overall conclusion from the findings is as follows. Although the figures in Table 2, are conservative we believe that they for most practical situations lead to the conclusion that the bias is negligible for all admissible sample sizes and, hence, that Pareto πps can safely be employed.

Table 2: Sample sizes that imply negligible bias

*Size Pattern category	Tolerance limit for negligibility								
	2%			1%			0.5 %		
	N	α	n_0	N	α	n_0	N	α	n_0
A	≥ 25	1	1	≥ 40	1	1	≥ 80	1	1
				$[25-40)$	1	3	$[50-80)$	1	3
							$[40-50)$	1	4
B	≥ 80	1	1	≥ 80	1	1	≥ 125	1	1
	$[25-80)$	1	2	$[40-80)$	1	3	$[100-125)$	1	3
							$[80-100)$	1	4
C	≥ 100	1	1	≥ 125	1	1	≥ 175	1	1
	≥ 40	0.8	1	≥ 100	0.9	1	≥ 150	0.9	1
	≥ 80	0.9	1	≥ 80	0.8	1	≥ 100	0.8	1
	≥ 25	0.5	1	≥ 40	0.5	1	≥ 80	0.5	1
D	≥ 150	0.9	1	≥ 175	0.8	1	≥ 125	0.5	1
	≥ 125	0.8	1	≥ 80	0.5	1	$[100-125)$	0.5	3
	≥ 80	0.5	1						
	$[50-80)$	0.5	2						

* See Table 1.
N is the population size and α states that the sample size should not exceed $\mathbf{n_{m,\alpha}}$ in (4.4). n_0 specifies a sufficiently large sample size, under the α-restriction, for negligible bias with specified tolerance. The study variable is presumed to be non-negative. Values for categories A and B are based on the weak quasi envelope (4.16), those for categories C and D on the (general) quasi envelope (4.7).

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**, pp. 457-469.

Aires, N. (2000). Comparisons between Conditional Poisson Sampling and Pareto πps Sampling Designs, *Journal of Statistical Planning and Inference*, **82**, No. 2, pp. 1-15.

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

Rosén, B. (2000). On Inclusion Probabilities of Order Sampling. To appear in *Journal of Statistical Planning and Inference.*

Saavedra (1995). Fixed Sample Size PPS Aproximations with a Permanent Random Number. *1995 Joint Statistical Meetings American Statistical Association.* Orlando, Florida.

Appendix 1

Appendix 2