R&D Report
Research – Methods – Development
2001:2

Statistiska centralbyrån Statistics Sweden

Order πps Inclusion Probabilities Are Asymptotically correct

Bengt Rosén

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 2001:2. Order πps inclusion probabilities are asymptotically correct / Bengt Rosén.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-2002-X101OP0102

R&D Report **2001:2**
Research – Methods – Development

Order πps Inclusion Probabilities Are Asymptotically correct

Bengt Rosén

R&D Report 2001:2
Research - Methods - Development

Order πps Inclusion Probabilities Are Asymptotically correct

Från trycket Augusti 2001
Producent Statistiska centralbyrån, *Statistics Sweden,* metodenheten
Box 24300, *SE*-104 51 STOCKHOLM

Förfrågningar Bengt Rosén, Uppsala universitet
bengt@math.uu.se
telefon 018- 471 32 23

ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 2001

June 2000

Order πps Inclusion Probabilities Are Asymptotically correct

Bengt Rosén

ABSTRACT

A particular class of sampling schemes with inclusion probability proportional to size (πps) was introduced in Rosén (1997), called order πps schemes. They were derived by limit considerations, and as a consequence their πps property is slightly approximate for finite samples. Rosén (2000 a) showed that the following holds under general conditions for three particular order πps scheme of special practical interest, Pareto, uniform and exponential order πps.

With $\lambda_k(n)$ and $\pi_k(n)$ for desired respectively factual inclusion probabilities for population unit k when the sample size is n ;

$\pi_k(n)/\lambda_k(n) \to 1$, uniformly over k as $n \to \infty$.

This entails that the schemes have asymptotically correct (= desired) inclusion probabilities. Here is shown that, as conjectured in Rosén (2000 a), the result holds not only for the mentioned particular schemes, but for order πps schemes very generally.

CONTENTS

Order πps Inclusion Probabilities Are Asymptotically Correct

1 Introduction and outline

A probability sample without replacement is to be drawn from the population $U = (1,2,...,N)$, using a sampling frame which one - to - one corresponds with the units in U. Moreover, the frame is presumed to contain values of a ***size variable***, $\mathbf{s} = (s_1, s_2, \ldots, s_N)$, $s_k > 0$. As is well known, if the size variable is fairly proportional to the (chief) study variable, estimation precision benefits from using a ***πps scheme***, i.e. a sampling scheme with ***sample inclusion probabilities*** $\pi_1, \pi_2, \ldots, \pi_N$ such that;

$$\pi_k \text{ is proportional to } s_k, \quad k = 1,2,...,N. \tag{1.1}$$

It is generally desirable that sampling schemes have predermined / fixed sample size. Accordingly, in the following we confine to πps schemes with fixed sample size n. Then, under (1.1) the ***desired inclusion probabilities*** $\lambda_1, \lambda_2, \ldots, \lambda_N$ are;

$$\lambda_k = n \cdot s_k / \sum_{j=1}^{N} s_j, \quad k = 1,2,\ldots,N. \tag{1.2}$$

This formula may yield λ:s which exceed 1, which is incompatible with being probabilities. If so, some adjustment has to be made, usually by introducing a "take for certain" stratum. In the sequel is presumed that adjustment already is made, so that $\lambda_k < 1$ holds for $k = 1,2,...,N$.

A "perfect" πps scheme satisfies (1.1) with π_k and λ_k being exactly equal for all k. In the following we are a bit more "generous". A sampling scheme which satisfies (1.3) below is accepted as a ***πps scheme (in wide sense)***;

$$\pi_k \approx \lambda_k \quad \textit{holds with good approximation for } k = 1,2,...,N. \tag{1.3}$$

The literature offers a multitude of πps schemes, among these the so called ***order πps schemes*** introduced in Rosén (1997). Such a scheme is specified by parameters (N, **s**, H, n), referred to as an (order πps) ***sampling situation***, with the following interpretation. N is the population size, $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ are size values, H(t) is the probability distribution function for a distribution on $[0, \infty)$ with density, and n is a predetermined sample size. The definition of an order πps scheme is stated below.

DEFINITION 1.1: An ***order πps*** sample from $U = (1,2,...,N)$, with ***size values*** $\mathbf{s} = (s_1, s_2, \ldots, s_N)$, ***shape distribution*** H(t) and ***sample size*** n is drawn as follows.

Step 1: Compute ***desired inclusion probabilities*** $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_N)$ by (1.2).

Step 2: Realize independent random variables $R_1, R_2, ..., R_N$ with uniform distri butions on $[0,1]$, and compute ***ranking variables*** Q as follows, where H^{-1} stands for inverse function;

$$Q_k = H^{-1}(R_k)/H^{-1}(\lambda_k), \quad k = 1,2,...,N. \tag{1.4}$$

Step 3: Finally, ***the sample*** consists of the units with the ***n smallest Q-values***.

It is by no means obvious that order πps schemes have the (wide sense) πps property (1.3), as indicated by their name. The chief aim in this paper is to prove that this in fact is true. More specifically we prove that (1.5) below holds under very general conditions. In (1.5), and henceforth, notations like $\pi_k(n)$ and $\lambda_k(n)$ indicate dependence on the sample size n.

$$\max_{k \in U} \left| \pi_k(n)/\lambda_k(n) - 1 \right| \to 0, \quad \text{as } n \to \infty. \tag{1.5}$$

Rosén (2000 a) showed that (1.5) holds for the three order πps schemes of greatest practical interest, which are specified by their shape distributions in (1.6) - (1.8) below. There the naming rule is that an order πps scheme is christened by the name of its shape distributions.

***Uniform order πps*:**	$H(t) = \min(t,1),\ 0 \le t < \infty,$ having $H^{-1}(\lambda) = \lambda.$	(1.6)
***Exponential order πps*:**	$H(t) = 1 - e^{-t},\ 0 \le t < \infty,$ having $H^{-1}(\lambda) = -\log(1-\lambda).$	(1.7)
***Pareto (order) πps*:**	$H(t) = t/(1-t),\ 0 \le t < \infty,$ having $H^{-1}(\lambda) = \lambda/(1-\lambda).$	(1.8)

Uniform order πps was introduced by Ohlsson (1990, 1998), who calls it ***sequential Poisson sampling***. The author and P. Saavedra (1995) independently came across Pareto order πps. Saavedra calls it ***odds ratio sequential Poisson sampling***. A "user's guide" for Pareto πps is presented in Rosén (2000 b). Aires (1999), Aires & Rosén (2000) and Rosén (2000 a) present findings from numerical investigations, which show that in particular for Pareto πps the approximation (1.3) works very accurately already for quite small sample sizes.

The concern in this paper is, however, not the small sample behavior of order πps inclusion probabilities but their *large sample behavior* for *general* shape distributions, not only for those in (1.6) - (1.8). As already stated, the chief result is that *(1.5) holds for a very wide class of shape distributions*.

Thereby we are addressing a result with simple formulation and general scope. We see it as a challenge that such a result ought to have a simple proof. This cannot be said about the proof in Rosén (2000 a), although it was confined to the three particular schemes. Even if proof ideas in this paper basically are the same as as in Rosén (2000 a), a considerably simpler proof a for a considerably more general result is presented. Matters are better understand by now, and we feel a bit embarrassed for earlier "clumsiness". We leave to the reader the challenge that the proof maybe can be even further simplifieid.

Throughout the paper P and E have their usual probability theory meanings, probability and expectated value. Moreover, log stands for natural logarithm and e = 2.718...

2 The chief result

In the subsequent limit considerations we work in the usual framework for finite population asymptotics: A ***sequence of populations*** U_q with sizes N_q, $q = 1,2,3,..$, such that;

$$N_q \to \infty, \quad \text{as } q \to \infty. \tag{2.1}$$

An index q signifies that the quantity relates to the sample from the q:th population. In particular, an order πps sampling situation is specified by parameters $(N_q, \mathbf{s}^{(q)}, H, n_q)$. Note that the shape distribution H is presumed to be the same in all situations. However, from now on we take a more general approach. Instead of viewing sizes $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ as primary parameters, that role is given to desired inclusion probabilities $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ., \lambda_N)$. They are presumed to be apriori specified, and their values may emanate from (1.2) or from somewhere else. As is well known, for a sampling scheme with fixed sample size, the inclusion probabilities add up to the sample size n. Accordingly, $\boldsymbol{\lambda}$ is presumed to satisfy;

$$0 < \lambda_k < 1,\ k = 1,2,\ldots,N, \quad \text{and} \quad \lambda_1 + \lambda_2 + \ldots + \lambda_N = n. \tag{2.2}$$

Hence, the q:th sampling situation is specified by parameters $(N_q, \boldsymbol{\lambda}^{(q)}, H, n_q)$, and the sampling scheme is defined by Steps 2 and 3 in Definition 1.1. Although size values no longer are explicitly involved, we continue to call the schemes "order πps schemes", even if ***schemes with varying inclusion probabilities*** would be more adequate. As before, $\pi_1, \pi_2, \ldots, \pi_N$ denote the true inclusion probabilities, in contrast to the desired ones, $\lambda_1, \lambda_2, \ldots, \lambda_N$. The task is to show that the approximation $\pi_k \approx \lambda_k$ under very general conditions works as stated in (1.5).

Some conditions on the parameters $(N_q, \lambda^{(q)}, H, n_q)$ must be imposed, though, and such conditions are introduced next.

The shape distribution H is a probability distribution on $[0, \infty)$ with density, denoted by h(t).

The ***support interval*** $[0, \tau_U)$ for H is defined by: $\tau_U = \sup\{t : H(t) < 1\}$. (2.3)

The density h(t) is continuous and strictly positive on $[0, \tau_U)$. (2.4)

Theorem 2.1 below concerns situations where the regularity condition (2.4) is met. It covers all order πps schemes which, as we understand it, may be considered for practical use, in particular those in (1.6) - (1.8). From a theoretical point of view, though, it is of interest to find out under how general conditions on H that (1.5) holds. Section 4 discusses ramifications of Theorem 2.1, i.a. some weakenings of the assumption (2.4).

Under (2.4) H(t) increases strictly and continuously on $[0, \tau_U)$, from 0 to 1. Hence, the ***inverse function*** $H^{-1}(\lambda)$, $0 < \lambda < 1$, can be, and is, defined in the simple and natural way: $H^{-1}(\lambda)$ = the (unique) t for which $H(t) = \lambda$. Boundary values for H^{-1} are set to $H^{-1}(0) = 0$, resp. $H^{-1}(1) = \tau_U$.

Next some notation relating to the desired inclusion probabilities;

$$\lambda^{(q)}_{min} = \min\{\lambda^{(q)}_1, \lambda^{(q)}_2, ..., \lambda^{(q)}_{N_q}\}, \quad \lambda^{(q)}_{max} = \max\{\lambda^{(q)}_1, \lambda^{(q)}_2, ..., \lambda^{(q)}_{N_q}\}. \tag{2.5}$$

We are now prepared to formulate the chief result.

THEOREM 2.1: A sequence $(N_q, \lambda^{(q)}, H, n_q)$, $q = 1, 2, 3, \ldots$, of order πps sampling situations is considered. The shape distribution H is presumed to satisfy (2.4). Then the following holds if conditions (i) - (iii) below are met;

$$\max_{k \in U_q} |\pi^{(q)}_k(n_q)/\lambda^{(q)}_k(n_q) - 1| \to 0, \text{ as } q \to \infty. \tag{2.6}$$

Conditions:

(i) $n_q \to \infty$, as $q \to \infty$, (2.7)

(ii) $\overline{\lim}_{q \to \infty} \lambda^{(q)}_{max} < 1$, (2.8)

(iii) $\overline{\lim}_{q \to \infty} \log(1/\lambda^{(q)}_{min})/\sqrt{n_q} = 0$. (2.9)

Remark: Condition (2.9) says that no λ-value is allowed to be "extremely small". Since (2.6) concerns the *relative* error in the approximation $\pi_k \approx \lambda_k$, a condition of this type is not surprising. We do not know, though, if (2.9) is a "precisely right" condition. ¤

The corollary below presents a version of the above result for the "traditional" πps situation, i.e. in terms of conditions on size values. In analogy with (2.5) we set;

$$s^{(q)}_{min} = \min\{s^{(q)}_1, s^{(q)}_2, ..., s^{(q)}_{N_q}\}, \quad s^{(q)}_{max} = \max\{s^{(q)}_1, s^{(q)}_2, ..., s^{(q)}_{N_q}\}. \tag{2.10}$$

COROLLARY 2.1: A sequence $(N_q, \mathbf{s}^{(q)}, H, n_q)$, $q = 1, 2, 3, \ldots$, of "ordinary" order πps situations is considered. λ:s are defined by (1.2), and H is presumed to fulfill (2.4). Then, (2.6) holds if (2.7), (2.8) and the following condition are met;

(iv) $\limsup_{q \to \infty} s^{(q)}_{max}/s^{(q)}_{min} < \infty$, (2.11)

(v) $(\log N_q)/\sqrt{n_q} \to 0$, as $q \to 0$. (2.12)

Justification of the corollary: From (2.2) follows generally that $\lambda^{(q)}_{max} \geq n_q/N_q$. When the λ:s are determined by (1.2) we have $\lambda^{(q)}_{min} = \lambda^{(q)}_{max} \cdot (s^{(q)}_{min}/s^{(q)}_{max})$. These two relations readily yield that (2.7) and (2.12) imply (2.9). ¤

3 Proof of Theorem 2.1

Until further notice we forget about the sequence setting and omit the index q. In particular, the parameters which specify the sampling situation are denoted by (N, λ, H, n).

3.1 A heuristic argument

To give background for the stringent proof we present a heuristic reasoning which sheds light on the approximation $\pi_k \approx \lambda_k$. By (1.4) and the fact that the R_k are uniform on $[0,1]$;

$$P(Q_k \le 1) = P(H^{-1}(R_k) \le H^{-1}(\lambda_k)) = P(R_k \le \lambda_k) = \lambda_k, \quad k = 1,2,\ldots,N. \tag{3.1}$$

Hence, Q_k takes its value in the interval $[0,1]$ with probability λ_k. This readily yields;

The expected number of Q:s with values in $[0,1]$ is $\lambda_1+\lambda_2+\ldots+\lambda_N =$ by (2.2) $= n$. (3.2)

Unit k is sampled if and only if Q_k takes its value among the n smallest Q:s. In view of (3.2) this holds roughly iff Q_k takes its value in the interval $[0,1]$, which by (3.1) has probability λ_k. Hence, the inclusion probability π_k can be expected to lie close to λ_k, i.e. the approximation $\pi_k \approx \lambda_k$ can be expected to work well. This argumentation is a bit too loose, though, to be a proof, but it provides a basis for a stringent proof.

3.2 Main steps in the proof

Start of the proof of Theorem 2.1 : The interest concerns the probability π_k that population unit k is sampled. For notational simplicity we choose, without loss of generality, k to be N and, thus, consider π_N. Since the Q - variables have continuous distributions we need not bother about ties in the definition of the order statistic Z(n) below.

$$Z(n) = \text{the n:th smallest among } Q_1, Q_2, \ldots, Q_{N-1}. \tag{3.3}$$

The fact that unit N is sampled if and only if $Q_N < Z(n)$ implies the relation (3.4), where $F_{Z(n)}(z)$ denotes the distribution of Z(n). The right-most equality is justified as follows. Since Z(n) is a function of $Q_1, Q_2, \ldots, Q_{N-1}$, which are presumed to be independent of Q_N, Z(n) and Q_N are independent, which yields $P(Q_N \le z \,|\, Z(n) = z) = P(Q_N \le z)$.

$$\pi_N = P(Q_N \le Z(n)) = \int_0^\infty P(Q_N \le z \,|\, Z(n) = z)\, dF_{Z(n)}(z) = \int_0^\infty P(Q_N \le z)\, dF_{Z(n)}(z). \tag{3.4}$$

For notational convenience we introduce the following shifted version of Z(n);

$$Z(n)^* = Z(n) - 1. \tag{3.5}$$

Then (3.4) may be written;

$$\pi_N - \lambda_N = \int_{-1}^\infty [P(Q_N \le 1+u) - \lambda_N]\, dF_{Z(n)^*}(u). \tag{3.6}$$

By splitting the domain of integration in (3.6) into $\{|u| \le z\}$ and $\{|u| > z\}$ and by using the trivial estimate $|P(Q_N \le 1+u) - \lambda_N| \le 1$ we get, for any $z > 0$;

$$|\pi_N/\lambda_N - 1| \le \frac{1}{\lambda_N} \cdot \left| \int_{|u| \le z} (P(Q_N \le 1+u) - \lambda_N)\, dF_{Z(n)^*}(u) \right| + P(|Z(n)^*| > z)/\lambda_N. \tag{3.7}$$

In the first round we want to exhibit the broad lines in the proof. For that reason we employ the contents in the following two lemmas straight away, while their proofs are deferred until next section. For fruitful use of (3.7) we need results about $P(Q_k \le 1 + u)$ as function of u and about the tails in the distribution of Z(n)*. First we introduce some notation.

Lower and upper bound functions, m(t;f) and M(t;f), for a function f(t) on $[0, \tau)$ are defined and denoted as follows. Note that m(t;f) decreases and M(t;f) increases as t increases.

$$m(t;f) = \inf_{0 \le s \le t} f(s), \quad M(t;f) = \sup_{0 \le s \le t} f(s), \quad 0 \le t < \tau. \tag{3.8}$$

Set, where h as usual denotes the density of H;

$$\tau_{max} = H^{-1}(\lambda_{max}),\quad u^* = (\tau_U - \tau_{max})/(2\cdot\tau_{max}),$$
$$\delta = m(\tau_{max}+u^*;h)/M(\tau_U;h),\quad \rho = M(\tau_U;h)/m(\tau_{max};h). \tag{3.9}$$

LEMMA 3.1: With notation according to (3.9) the following holds under (2.4);

$$P(Q_k \le 1+u) = \lambda_k + u\cdot\lambda_k\cdot\eta_k(u),\quad -1 \le u, \tag{3.10}$$

with

$$\delta \le \eta_k(u) \le \rho,\quad -1 \le u < u^*. \tag{3.11}$$

LEMMA 3.2: With notation according to (3.5), (3.8) and (3.9) we have under (2.4);

$$P(|Z(n)^*| > z) \le 4.2\cdot\exp\{-z\cdot\sqrt{n}\cdot\delta/(2\cdot\sqrt{1+z\cdot\rho})\},\ 0 \le z \le u^*, \tag{3.12}$$

provided that the following condition is met;

$$n\cdot[1-(\lambda_{max}+\rho\cdot u^*+(u^*)^2\cdot\rho^2+1/4n)] \ge 1. \tag{3.13}$$

Continuation of the proof of Theorem 2.1: We now return to the sequence situation in Theorem 2.1, and pursue (3.7) in that setting. Then λ_{max}, τ_{max}, u^*, δ and ρ depend on q. However, as is readily realized, under (2.4) and (2.8) these quantities are uniformly bounded to the effect that there exist values λ_{max}, τ_{max}, u^*, δ and ρ such that;

$$\lambda_{max}^{(q)} \le \lambda_{max} < 1,\quad \tau_{max}^{(q)} \le \tau_{max} < \tau_U,\quad u^{(q)*} \ge u^* > 0,\quad \delta^{(q)} \ge \delta > 0,\quad \rho^{(q)} \le \rho < \infty,$$
$$q = 1,2,3,\ldots. \tag{3.14}$$

A consequence of (3.14) and (2.7) is that u^* can be, and is presumed to be, chosen so that for some $q_0 < \infty$ (3.13) holds for $q \ge q_0$. Moreover, z is presumed to satisfy;

$$0 < z \le u^*. \tag{3.15}$$

Then Lemmas 3.1 and 3.2 can be employed. By using (3.10) and (3.11) in the integral in (3.7);

$$|\pi_{N_q}^{(q)}/\lambda_{N_q}^{(q)} - 1| \le \rho\cdot\int_{|u|\le z}|u|\,dF_{Z_q(n_q)^*}(u) + P(Z_q(n_q)^* > z)/\lambda_N,\quad q_0 \le q. \tag{3.16}$$

By (3.12);

$$P(|Z_q(n_q)^*| > z) \le 4.2\cdot\exp\{-z\cdot\delta\cdot\sqrt{n_q}/(2\cdot\sqrt{1+z\cdot\rho})\},\ 0 < z \le u^*,\ q_0 \le q. \tag{3.17}$$

A straightforward consequence of (3.17) and (2.7) is;

$$Z_q(n)^* \text{ converges in probability to } 0,\ \text{as } q \to \infty. \tag{3.18}$$

From (3.18) follows that the integral in (3.16) tends to 0 as $q \to \infty$. Moreover, from (3.17) and (2.9) follows readily that $P(Z_q(n_q)^* > z)/\lambda_{N_q} \to 0$ as $q \to \infty$.

Thereby the convergence (2.6) is established for population unit N_q. However, N_q can be seen as an arbitrary unit in U_q, and the bounds used in the proof hold uniformly over the population (as well as over q). Hence, Theorem 2.1 is proved but for Lemmas 3.1 and 3.2. ¤

3.3 Proofs of Lemmas 3.1 and 3.2

3.3.1 Proof of Lemma 3.1

By (1.4) and the fact that R_k is uniformly distributed on [0,1];

$$P(Q_k \le 1+u) = P(H^{-1}(R_k) \le (1+u)\cdot H^{-1}(\lambda_k)) =$$
$$= P(R_k \le H((1+u)\cdot H^{-1}(\lambda_k)) = H(H^{-1}(\lambda_k) + u\cdot H^{-1}(\lambda_k)),\quad -1 < u < \infty. \tag{3.19}$$

We shall use the Taylor expansion $f(x+\Delta) = f(x) + \Delta\cdot f'(\xi)$, for some ξ between x and $x+\Delta$. By (2.4), H is differentiable on $[0,\tau_U)$ with continuous derivative $H'(t) = h(t)$. Hence, provided

that $H^{-1}(\lambda_k) + u \cdot H^{-1}(\lambda_k) \in [0, \tau_U)$, which is readily checked to hold for $-1 \le u \le u^*$, Taylor expansion of the last term in (3.19), with $x = H^{-1}(\lambda_k)$ and $\Delta = u \cdot H^{-1}(\lambda_k)$, leads to (3.20) below. Note that $H(H^{-1}(\lambda_k)) = \lambda_k$.

$$P(Q_k \le 1+u) = \lambda_k + u \cdot H^{-1}(\lambda_k) \cdot \vartheta_k(u), \tag{3.20}$$

where

$$\vartheta_k(u) = h(H^{-1}(\lambda_k) + \theta(u) \cdot u \cdot H^{-1}(\lambda_k)) \quad \text{for some } 0 < \theta(u) < 1. \tag{3.21}$$

From (3.21) follows readily, with m and M according to (3.8);

$$m(\tau_{max}+u^*; h) \le \vartheta_k(u) \le M(\tau_U; h). \tag{3.22}$$

By combining (3.20) and (3.21) with the estimate (3.23) below, Lemma 3.1 follows.

With m, M and τ_{max} as in (3.8) and (3.9):

$$\lambda_k / M(\tau_{max}; h) \le H^{-1}(\lambda_k) \le \lambda_k / m(\tau_{max}; h), \quad k = 1,2,\ldots,N. \tag{3.23}$$

To prove (3.23) we use the representation $H(t) = \int_0^t h(s)\,ds$, $0 \le t < \tau_U$ which readily yields;

$$t \cdot m(\tau_{max}; h) \le H(t) \le t \cdot M(\tau_{max}; h), \quad 0 < t \le \tau_{max}. \tag{3.24}$$

Inversion of (3.24) yields (3.23). ¤

3.3.2 Proof of Lemma 3.2

The task is to derive an estimate of $P(|Z(n)^*| > z)$. We have;

$$P(|Z(n)^*| > z) = P(|Z(n) - 1| > z) = P(Z(n) < 1 - z) + P(Z(n) > 1+z), \quad z > 0. \tag{3.25}$$

In the random variables introduced below, $\mathbf{1}(A)$ denotes the indicator of the event A. Since the Q_k are presumed to be positive we confine to the domain $-1 \le u < \infty$.

$$X(u)_k = \mathbf{1}(Q_k \le 1+u), \quad k = 1,2,\ldots,N, \quad -1 \le u < \infty, \tag{3.26}$$

$$S(u) = X(u)_1 + X(u)_2 + \ldots + X(u)_{N-1}, \quad -1 \le u < \infty. \tag{3.27}$$

It is readily checked that the following relations hold;

$$\{Z(n) \le 1+u\} = \{S(u) \ge n\}, \quad \{Z(n) > 1+u\} = \{S(u) \le n-1\}, \quad -1 \le u < \infty. \tag{3.28}$$

From (3.27) is seen that information about the tails of the distribution of Z(n) can be obtained via information about the distribution of S(u), which is a sum of independent Bernoulli variables. We insert a general result about such sums.

LEMMA 3.3: Let $S = X_1 + X_2 + \ldots + X_R$ be a sum of independent Bernoulli variables with $P(X_k = 1) = p_k$, $k = 1,2,\ldots,R$. Set;

$$\mu = \sum_{k=1}^{R} p_k \quad \text{and} \quad \sigma^2 = \sum_{k=1}^{R} p_k \cdot (1 - p_k). \tag{3.29}$$

Then, provided that $\sigma \ge 1$, the following holds;

$$P(S \ge \mu + \beta \cdot \sigma) \le 2.1 \cdot e^{-\beta}, \quad -\infty < \beta < \infty. \tag{3.30}$$

Alternative forms of the estimate in (3.30) are stated below.

$$\text{For any } b\text{: } P(S \ge b) \le 2.1 \cdot \exp\{-(b-\mu)/\sigma\}, \quad P(S \le b) \le 2.1 \cdot \exp\{-(\mu - b)/\sigma\}. \tag{3.31}$$

Justification : The estimate (3.30) is derived as Lemma 2.2 in Rosén (2000 a). The left hand part in (3.31) is a straightforward modification of (3.30). The right part is obtained e.g. from the left part with b exchanged for R - b and variables $X'_k = 1 - X_k$, $k = 1,2,,R$, which also are Bernoulli variables, with $S' = R - S$, $\mu' = R - \mu$ and $\sigma' = \sigma$. Hence Lemma 3.3 is justified. ¤

Continuation of the proof of Lemma 3.2 : The relations in (3.28) in combination with the bounds in (3.31) with b = n yield the following estimates, for $\mu(u)$ and $\sigma(u)$ in accordance with (3.29), provided that $\sigma(u) \geq 1$;

$$P(Z(n) > 1+u) = P(S(u) \leq n-1) \leq 2.1 \cdot \exp\{-[\mu(u) - (n-1)]/\sigma(u)\}, \quad 0 \leq u < \infty. \tag{3.32}$$

$$P(Z(n) \leq 1+u) = P(S(u) \geq n) \leq 2.1 \cdot \exp\{-[n-\mu(u)]/\sigma(u)\}, \quad -1 \leq u < 0, \tag{3.33}$$

To continue we need expressions for $\mu(u)$ and $\sigma(u)$ according to (3.29). Here R = N-1 and;

$$p_k(u) = P(X(u)_k = 1) = P(Q_k \leq 1+u) = \text{by (3.10)} = \lambda_k + u \cdot \lambda_k \cdot \eta_k(u), \quad -1 \leq u. \tag{3.34}$$

Hence;

$$\mu(u) = \sum_{k=1}^{N-1} p_k(u) = \sum_{k=1}^{N-1} \lambda_k + u \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u) \ = \text{by (2.2)} = n - \lambda_N + u \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u), \tag{3.35}$$

$$\sigma^2(u) = \sum_{k=1}^{N-1} p_k(u) \cdot [1 - p_k(u)] = \sum_{k=1}^{N} \lambda_k \cdot (1-\lambda_k) + \\ + u \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u) \cdot (1 - 2 \cdot \lambda_k) - u^2 \cdot \sum_{k=1}^{N-1} \lambda_k^2 \cdot \eta_k(u)^2 - \lambda_N \cdot (1-\lambda_N). \tag{3.36}$$

By (3.11) and (2.2) we get from (3.35) when $0 \leq u \leq u^*$ and $n \geq 2$;

$$\mu(u) - (n-1) = 1 - \lambda_N + u \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u) \ \geq u \cdot \delta \cdot (n - \lambda_N) \geq u \cdot n \cdot \delta / 2. \tag{3.37}$$

Analogously for $u < 0$ and $n \geq 2$;

$$n - \mu(u) = \lambda_N - u \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u) \geq |u| \cdot \delta \cdot \sum_{k=1}^{N-1} \lambda_k = |u| \cdot \delta \cdot (n - \lambda_N) \geq |u| \cdot n \cdot \delta / 2. \tag{3.38}$$

Next we consider $\sigma^2(u)$, and start with an upper bound. By (3.36), the estimate $-1 < 1 - 2 \cdot \lambda_k < 1$, $k = 1,2,\ldots,N$, (3.11) and (2.2) we have;

$$\sigma^2(u) \leq \sum_{k=1}^{N} \lambda_k \cdot (1-\lambda_k) + |u| \cdot \sum_{k=1}^{N} \lambda_k \cdot \eta_k(u) \leq \\ \leq n + |u| \cdot \rho \cdot \sum_{k=1}^{N} \lambda_k \leq n \cdot (1 + |u| \cdot \rho). \tag{3.39}$$

By employing the estimates (3.37) - (3.39) in (3.32) and (3.33) it is seen that $P(Z(n) < 1-z)$ and $P(Z(n) > 1+z)$ for $z > 0$ both are dominated by $2.1 \cdot \exp\{-|z| \cdot \sqrt{n} \cdot \delta / (2 \cdot \sqrt{1 + z \cdot \rho}\}$. This together with (3.25) yields (3.12).

However, we are not entirely through yet. Lemma 3.3 contains the premise $\sigma(u) \geq 1$, and it remains to formulate conditions for that to be fulfilled. Again we start from (3.36). The estimates $1 - \lambda_k > 1 - \lambda_{max}$, $-1 < 1 - 2 \cdot \lambda_k < 1$, and $\lambda_k \cdot (1 - \lambda_k) \leq 1/4$, $k = 1,2,\ldots,N$, and (3.11) yield;

$$\sigma^2(u) \geq \sum_{k=1}^{N} \lambda_k \cdot (1-\lambda_k) - |u| \cdot \sum_{k=1}^{N-1} \lambda_k \cdot \eta_k(u) \ - u^2 \cdot \sum_{k=1}^{N-1} \lambda_k^2 \cdot \eta_k(u)^2 - \lambda_N \cdot (1-\lambda_N) \geq \\ \geq n \cdot (1 - \lambda_{max}) - |u| \cdot \rho \cdot \sum_{k=1}^{N-1} \lambda_k - u^2 \cdot \rho^2 \cdot \sum_{k=1}^{N-1} \lambda_k - \lambda_N \cdot (1-\lambda_N) \geq \\ \geq n \cdot [1 - (\lambda_{max} + |u| \cdot \rho + u^2 \cdot \rho^2 + 1/4n)]. \tag{3.40}$$

From (3.40) is seen that (3.13) implies that $\sigma(u) \geq 1$. Thereby the lemma is proved. ¤

4 Ramifications of Theorem 2.1

4.1 On the rate of convergence in (2.6)

The estimates used in the proof of Theorem 2.1 allow derivation of bounds for the rate of convergence in (2.6). Below we present a sharpened version of Theorem 2.1.

THEOREM 4.1: With notation and assumptions as in Theorem 2.1;

$$\overline{\lim}_{q\to\infty}\sqrt{n_q}\cdot\max_{k\in U_q}|\pi_k^{(q)}(n_q)/\lambda_k^{(q)}(n_q)-1|<\infty. \tag{4.1}$$

Proof: We start from (3.16), and shall use the following general formula (which is readily shown by partial integration) for any non-negative random variable X with density;

$$\int_0^z x\,dF_X(x)=\int_0^z P(X>x)\,dx+z\cdot P(X>z),\quad 0\le z. \tag{4.2}$$

By employing (4.2), the integral in (3.16) can be transformed as follows;

$$\int_{|u|\le z}|u|\,dF_{Z(n_q)^*}(u)=\int_0^z P(|Z(n_q)^*|>u)\,du+z\cdot P(|Z(n_q)^*|>z). \tag{4.3}$$

Combination of (3.16), (4.3) and (3.17) yields;

$$|\pi_{N_q}^{(q)}/\lambda_{N_q}^{(q)}-1|\le$$

$$\le 4.2\cdot\left[\rho\cdot\int_0^\infty e^{-u\cdot\sqrt{n_q}\cdot\delta/(2\cdot\sqrt{1+\rho\cdot u})}du+(\rho\cdot z+1/\lambda_{N_q})\cdot e^{-z\cdot\sqrt{n_q}\cdot\delta/(2\cdot\sqrt{1+\rho\cdot z})}\right]. \tag{4.4}$$

Next we employ the following two general estimates, which are proved below.

For $\alpha>0$ and $\rho>0$: $\int_0^\infty e^{-\alpha\cdot u/\sqrt{1+\rho\cdot u}}\,du\le\sqrt{2}/\alpha+4\cdot\rho/\alpha^2$. (4.5)

For $\alpha>0$ and $0\le z\le 1$: $z\cdot e^{-z\cdot\alpha/\sqrt{1+\rho\cdot z}}\le\sqrt{1+\rho}/(\alpha\cdot e)$. (4.6)

By using (4.5) and (4.6) in (4.4), with $\alpha=\sqrt{n_q}\cdot\delta/2$;

$$|\pi_{N_q}^{(q)}/\lambda_{N_q}^{(q)}-1|\le 4.2\cdot\left[2\cdot\sqrt{2}\cdot\rho/(\delta\cdot\sqrt{n_q})+16\cdot\rho/(\delta^2\cdot n_q)+\right.$$

$$\left.+2\cdot e^{-1}\cdot\rho\cdot\sqrt{1+\rho}/(\delta\cdot\sqrt{n_q})+e^{-z\cdot\sqrt{n_q}\cdot\delta/(2\cdot\sqrt{1+\rho\cdot z})}/\lambda_{N_q}\right]. \tag{4.7}$$

The claim in (4.1) now follows readily from (4.7) and (2.9).

Hence, it only remains to prove (4.4) and (4.5). The latter follows from the straightforward inequality $z\cdot\exp\{-\beta\cdot z\}\le 1/(\beta\cdot e)$ together with $1+\rho\cdot z\le 1+\rho$, $\rho>0$ and $0\le z\le 1$. The inequality (4.4) can be shown as follows.

$$\alpha\cdot\int_0^\infty e^{-\alpha\cdot u/\sqrt{1+\rho\cdot u}}\,du=[\alpha\cdot u=s]=\int_0^\infty e^{-s/\sqrt{1+s\cdot\rho/\alpha}}\,ds\le\int_0^{\alpha/\rho}e^{-s/\sqrt{2}}ds+\int_{\alpha/\rho}^\infty e^{-s/\sqrt{2\cdot s\cdot\rho/\alpha}}\,ds\le$$

$$\le\int_0^\infty e^{-s/\sqrt{2}}\,ds+\int_0^\infty e^{-\sqrt{s}/\sqrt{2\cdot\rho/\alpha}}\,ds=[\sqrt{s}/\sqrt{2\cdot\rho/\alpha}=x]=\sqrt{2}+4\cdot\frac{\rho}{\alpha}\cdot\int_0^\infty x\cdot e^{-x}\,dx=\sqrt{2}+4\cdot\rho/\alpha. \tag{4.8}$$

4.2 Comments on weakenings of condition (2.4)

Theorem 2.1 is proved under the regularity assumption on the shape distribution H which is stated in (2.4). A natural question is if (2.6) still holds if (2.4) is not satisfied. Upon some thought it is realized that the previous proof can be modified in a straightforward manner to work also under the following weaker assumptions on H.

(2.4) is changed to: h(t) is piece-wise continuous and strictly positive on $[0,\tau_U)$. (4.9)

(2.4) is changed to: h(t) is piece-wise continuous and strictly positive on $(0, \tau_U)$, together with the assumption $\underline{\lim}_{q\to\infty} \lambda_{\min}^{(q)} > 0$. (4.10)

However, it is an open question what happens to (2.6) generally if h(t) is not strictly positive on $[0, \tau_U)$, in particular if $h(t) \equiv 0$ in a whole vicinity of $t = 0$.

References

Aires, N. & Rosén, B. (2000). On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling. Statistics Sweden R&D Report 2000:2.

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. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, **14**, 149-162.

Rosén, B. (1997). On Sampling with Probability Proportional to Size. *J Stat. Plan. Inf.*, **62** 159 - 191.

Rosén, B. (2000 a). On Inclusion Probabilities for Order Sampling. *J Stat. Plan. Inf.*, **90** 117-143.

Rosén, B. (2000 b). A User's guide to Pareto ps Sampling. In *Proceedings from the Second International Conference on Establishment Surveys, Buffalo 2000.*

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

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och Metodenheten gemensam publikationsserie, som 1988-01-01 ersatte de tidigare "gula" och "gröna" serierna. I serien ingick fram till årsskiftet 1992-93 även **Abstracts** (sammanfattning av metodrapporter från SCB).

Reports published during 1999 and onwards:

1999:1 Täckningsproblem i Registret över totalbefolkning RTB. Skattning av övertäckning med en indirekt metod *(Jan Qvist)*

1999:2 Bortfallsbarometer nr 14 *(Per Nilsson, Antti Ahtiainen, Mats Bergdahl, Tomas Garås, Jan Qvist och Charlotte Strömstedt)*

1999:3 Att mäta statistikens kvalitet *(Claes Andersson, Håkan L. Lindström och Thomas Polfeldt)*

2000:1 Kalibrering av vikter – beskrivning av tekniken och de SCB-fall den prövats i *(Sixten Lundström et al)*

2000:2 On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling *(Nibia Aires and Bengt Rosén)*

2000:3 Bortfallsbarometer nr 15 *(Per Nilsson, Ann-Louise Engstrand, Sara Tångdahl, Stefan Berg, Tomas Garås och Arne Holmqvist)*

2000:4 Bortfallsanalys av SCB-undersökningarna HINK och ULF *(Jan Qvist)*

2000:5 Generalized Regression Estimation and Pareto πps *(Bengt Rosén)*

2000:6 A User's Guide to Pareto πps Sampling *(Bengt Rosén)*

2001:1 Det statistiska registersystemet. Utvecklingsmöjligheter och förslag *(SCB, Registerprojektet)*

2001:2 Order πps Inclusion Probabilities Are Asymptotically correct *(Bengt Rosén)*

ISSN 0283-8680

Tidigare utgivna **R&D Reports** kan beställas genom Katarina Klingberg, SCB, MET, Box 24 300, 104 51 STOCKHOLM (telefon 08-506 942 82, fax 08-506 945 99, e-post katarina.klingberg@scb.se).
R&D Reports from 1988-1998 can - in case they are still in stock - be ordered from Statistics Sweden, attn. Katarina Klingberg, MET, Box 24 300, SE-104 51 STOCKHOLM (telephone +46 8 506 942 82, fax +46 8 506 945 99, e-mail katarina.klingberg@scb.se).

SCB-Tryck, Örebro 2001.08