# On Maximum Entropy $\pi$ps-sampling With Fixed Sample Size

JOHAN JONASSON AND OLLE NERMAN *
Department of Mathematics
Chalmers University of Technology
S-412 96 Göteborg
Sweden

September 1, 1997

**Abstract**

A $\pi$ps-sample is a sample, $s$, chosen from the population $\{1, \ldots, N\}$ under the condition that each individual, $i$, has a predetermined probability, $\pi_i$ of being included in $s$. The case where the sample size, $|s|$, is fixed is considered. An alternative proof is given of Hajek's result that if $s$ is chosen by ordinary Poisson sampling with inclusion probabilities $p_i$, then the conditional distribution of $s$ given $|s| = n$ has maximum entropy under the resulting inclusion probabilities, $P(i \in s \,||s| = n)$. It is also shown that the $p_i$'s can be chosen in such a way that for each $i = 1, \ldots, N$, $P(i \in s \,||s| = n) = \pi_i$ and $\sum_{i=1}^{N} p_i = n$. Asymptotic normality of Horvitz-Thompson estimators of the population total for such conditional Poisson sampling procedures is shown under quite general conditions using stochastic monotonicity arguments. The question of how the $p_i$'s and the $\pi_i$'s relate for large samples is treated and it is shown that in most, but not all, senses these are asymptotically exchangeable. Some consideration is also given to the computational problem of choosing a conditional Poisson sample in practice.

**Keywords:** $\pi$ps-sampling, Poisson sampling, central limit theory, entropy.

## 1 Introduction

In sampling theory the situation is often the following. In a population consisting of $N$ individuals (or companies, cities, schools, etc) labelled $1, \ldots, N$, each individual, $i$, has a weight $x_i$ (such as income, profit, number of inhabitants, number of students, etc.). We are interested in the *population total*, $t_x = \sum_{i=1}^{N} x_i$ (alternatively, the *population mean*, $t_x/N$). Since it is usually practically impossible to measure the weight of every individual, one chooses at random a sample, $s$, from the population

*Supported by the Bank of Sweden, Tercentenary Foundation Grant no. 93:039:03.

and uses this sample to estimate $t_x$. In the simplest situation, where all the individuals have the same probabilities of being included in $s$, it is intuitively obvious that the best estimator of $t_x$ is

$$\frac{N}{|s|}\sum_{i\in s} x_i$$

where $|s|$ is the size of the sample. This situation has been thoroughly analyzed and it can be found in any good book on sampling theory.

In many situations however, one has access to some auxiliary information about the individuals in the following way. For each individual a quantity $y_i$ is known and it is believed that $y_i$ is approximately proportional to the weight $x_i$. (This is for instance the case if the $i$'s are schools, the $x_i$'s are the number of students and the $y_i$'s are the number of teachers). In order to use this information to get better estimations of $t_x$ it is natural to make the choice of the sample in such a way that the inclusion probabilities for different individuals are different. To be exact, one tries to make the choice so that the inclusion probability, $\pi_i$, of $i$ is proportional to $y_i$. To estimate $t_x$ one uses the *Horvitz-Thompson esimator* introduced by Horvitz and Thompson (1952):

$$\hat{t}_x = \sum_{i=1}^{N} I_i \frac{x_i}{\pi_i}$$

where $I_i$ is the indicator $I_{\{i\in s\}}$. The Horvitz-Thompson estimator is clearly unbiased and if the correlation between the $x_i$'s and the $y_i$'s is strong, it has very low variance. To get a feeling for this, consider the extreme case when $y_i$ is exactly proportional to $x_i$. Then $\hat{t}_x = c|s|$ for some constant $c$ so that $\mathbf{Var}(\hat{t}_x) = c^2\mathbf{Var}(|s|)$ and if the sample has fixed size, $n$, the variance of $\hat{t}_x$ is zero.

Sampling with different probabilities for different individuals as described above is called *$\pi$ps-sampling* (or sometimes pps-sampling, though this term is usually reserved for sampling with replacement), where $\pi$ps stands for "inclusion probabilities proportional to size". As we just have seen, $\pi$ps-sampling gives low variance for the Horvitz-Thompson estimator if the $\pi_i$'s can be chosen to be approximately proportional to the weights $x_i$, a fact which of course makes it very attractive. However, one problem with $\pi$ps-sampling is that it is not at all obvious what is the best way to choose the sample. In the case where all the inclusion probabilities are equal, it is intuitively clear that if you want a fixed sample size, $n$, you should choose $s$ uniformly in the set of all possible samples of size $n$ and if you allow random sample size you should choose the typical individual with probability equal to the inclusion probability independently of the others. One sense in which these methods are characterized is that they maximize the *entropy* of $s$ in these situations, i.e. they maximize the function

$$-\sum_{s_0} P(s = s_0)\log P(s = s_0)$$

where the sum is over all samples of size $n$ in the first case and over all subsets of the population in the second. In a general sampling situation we want to find a sampling scheme which gives a distribution of $s$ satisfying a set of conditions (such as fixed sample size, inclusion probabilities, pairwise inclusion probabilities, etc). Since the distribution with maximal entropy in a class of distributions is the "most random" distribution, or if you like, the "most informative" distribution in that

class, it seems like a reasonable criterion for a best sampling scheme that it should maximize the entropy of $s$ in the class of distributions satisfying these conditions.

In Section 2 we establish such a sampling scheme; conditional Poisson sampling. Once this has been done, the next natural step to take is to see if it is useful in practice. One obvious first step in this direction is to try to answer questions about the asymptotic behavior of the Horvitz-Thompson estimator for large samples from large populations. There are earlier central limit theorems for Horvitz-Thompson estimators in related sampling procedures, e.g. in Rosèn (1995) and Olsson (1995) where so called sequential sampling procedures are treated. Is there a central limit theorem in the conditional Poisson sampling situation? The answer is yes and this question is treated in Sections 3-7.

In Section 8, we consider the question of how the $p_i$'s and the $\pi_i$'s relate for large samples, in particular if they are asymptotically exchangeable.

In the last section the computational problem of generating a Poisson sample in practice is treated. We show that by using a more sophisticated device than the straightforward method of repeatedly choosing unconditional Poisson samples and rejecting them until exactly $n$ indivuiduals are chosen, the number of operations can be cut down considerably.

# 2 The Maximum Entropy Method

Let us go back to the $\pi$ps-sampling situation. If you allow random sample size the sampling scheme in question is fairly easy to find; pick the individuals with probability $\pi_i$ independently of the others. This method was introduced by Hajek (1964). It is known as *Poisson sampling* and it has maximum entropy in the class of all distributions on $s$ satisfying $P(i \in s) = \pi_i$ for every $i$. In the case of a fixed sample size, $n$, however, the problem is harder. A great number of different $\pi$ps-sampling procedures with fixed sample size yielding the right inclusion probabilities have been developed through the years. Brewer and Hanif (1983) give no less than 50 $\pi$ps-sampling procedures, most of them with fixed sample size, but none of these give maximum entropy for the distribution of $s$.

Hajek (1981) shows that conditioning on the sample size in Poisson sampling yields the maximum entropy distribution of $s$ under the condition of fixed sample size and under the resulting conditional inclusion probabilities. Theorem 2.1, which has a very short and simple proof, implies Hajek's result. The problem is that if we choose a Poisson sample with inclusion probabilities $p_i$, then it is almost never true that $P(i \in s \,||s| = n) = p_i$. Since we would like to have $P(i \in s \,||s| = n) = \pi_i$ it is natural to ask if the $p_i$'s can be chosen in such a way that this holds true. In Theorem 2.2 it is shown that such a choice can be made, yielding the existence of a procedure for maximum entropy $\pi$ps-sampling with fixed sample size and the desired inclusion probabilities.

THEOREM 2.1 *Let $X$ be an $\Omega$-valued random object where $\Omega$ is a finite set and let $Y$ be a discrete $\sigma(X)$-mesurable random variable. Assume that $A_1, \ldots, A_k$ are subsets of $\Omega$ and that the entropy, $H(P_X) = H(X)$, of the distribution of $X$ is maximal in the class of distributions, $P$, on $\Omega$ satisfying $P(A_i) = a_i$, $i = 1, \ldots, k$. Then it is true for every $y$ in the support of $Y$ that the conditional entropy, $H(X|Y = y)$, is maximal in the class of distributions, $Q$, on $\{Y = y\}$ satisfying $Q(A_i \cap \{Y = y\}) = P_X(A_i|Y = y)$, $i = 1, \ldots, k$.*

*Proof.* Use the standard formula

$$H(X,Y) = H(Y) + \sum_y P(Y=y)H(X|Y=y)$$

and note that since $Y$ is $\sigma(X)$-measurable, $H(X,Y) = H(X)$. If $H(X|Y=y_0)$ is not maximal for some $y_0$ in the support of $Y$, then we can change the distribution of $X$ on $\{Y=y_0\}$ to increase $H(X|Y=y_0)$. Since none of the other expressions in the above formula changes this means that $H(X)$ also increases, a contradiction. □

Letting $\Omega$ be the family of subsets of $\{1,\dots,N\}$, $X=s$, $Y=I_{\{|s|=n\}}$ and $A_i=\{i\in s\}$ in the above theorem yields Hajek's result. This, however, only says that conditional Poisson sampling yields the maximum entropy distribution for the inclusion probabilities given by $P(i\in s\,||s|=n)$; it is not à priori clear that the $p_i$'s can be chosen so that $P(i\in s\,||s|=n)=\pi_i$ for any given set of $\pi_i$'s satisfying $\sum_{i=1}^N \pi_i = n$. For that we would like to solve the following system of equations:

$$\frac{\sum_{A\in\mathcal{A}_n(i)}(\prod_{j\in A}p_j\prod_{k\notin A}(1-p_k))}{\sum_{A\in\mathcal{A}_n}(\prod_{j\in A}p_j\prod_{k\notin A}(1-p_k))}=\pi_i, i=1,\dots,k \tag{1}$$

where $\mathcal{A}_n$ is the class of all possible samples of size $n$ and $\mathcal{A}_n(i)$ is the set of samples of $\mathcal{A}_n$ containing $i$. The following theorem states that this system of equations always has a unique solution such that $\sum_{i=1}^N p_i = n$ which is desirable from a practical point of view.

THEOREM 2.2 *For any set of values of $\pi_1,\dots,\pi_N$ such that $0<\pi_i<1$ and $\sum_{i=1}^N\pi_i=n$, the system of equations (1) has a unique solution such that $\sum_{i=1}^N p_i=n$.*

*Proof.* For simplicity of notation, rewrite (1) as

$$\pi_1 = f_1(p_1,\dots,p_N) \tag{e1}$$

$$\vdots$$

$$\pi_N = f_N(p_1,\dots,p_N) \tag{eN}$$

where

$$f_i(p_1,\dots,p_N)=\frac{\sum_{A\in\mathcal{A}_n(i)}\prod_{j\in A}p_j\prod_{k\notin A}(1-p_k)}{\sum_{A\in\mathcal{A}_n}\prod_{j\in A}p_j\prod_{k\notin A}(1-p_k)}=\frac{\sum_{A\in\mathcal{A}_n(i)}\prod_{j\in A}p_j/(1-p_j)}{\sum_{A\in\mathcal{A}_n}\prod_{j\in A}p_j/(1-p_j)}.$$

Note that all the $f_i$'s are continuous and that $f_i$ is (strictly) increasing as a function of $p_i$ and decreasing as a function of $p_j$ for any specific $j\neq i$. The first of these monotonicities is obvious whereas the others are more subtle. To see the latter, first note that $f_i(p_1,\dots,p_n)$ can, if seen as a function $h_i$ of $p_j$, be written as

$$h_i(p_j)=\frac{ap_j/(1-p_j)+b}{cp_j/(1-p_j)+d}$$

for nonnegative constants $a$, $b$, $c$ and $d$. Therefore $h_i(p_j)$ is either decreasing, increasing or constant. Thus $\lim_{p_j\to 0}h_i(p_j)$ and $\lim_{p_j\to 1}h_i(p_j)$ exist. The first of these limits is the inclusion probability for $i$ in a Poisson sample from $\{1,2,\dots,j-1,j+1,\dots,N\}$ given that the sample size is $n$ whereas the second is the corresponding probability

given that the sample size is $n-1$. Therefore Proposition 6.2 below yields that $h_i(p_j)$ is in fact decreasing.

We are going to solve the system of equations by calibrating the $p_i$'s one by one keeping the remaining ones fixed. In doing so we assume without loss of generality that $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_N$. Note that it is always automatically true that $\sum_{i=1}^N f_i(p_1,\ldots,p_N) = n$ for any set of values of the $p_i$'s and that $f_1,\ldots,f_N$ are ordered in the same order as $p_1,\ldots,p_N$ are.

We start the procedure by fixing $p_2,\ldots,p_N$ at arbitrary values in $(0,1)$. Since $f_1(p_1,\ldots,p_N)$ is increasing in $p_1$ and tends to 0 as $p_1 \to 0$ and to 1 as $p_1 \to 1$ there is by continuity a unique $p_1$ satisfying (e1). This $p_1$ can be written as a function $g_1^1(p_2,\ldots,p_N)$ where $g_1^1$ is of course continuous. Inserting this into (e2) yields the equation

$$\pi_2 = f_2(g_1^1(p_2,\ldots,p_N),p_2,\ldots,p_N),$$

which we next, for arbitrary $p_3,\ldots,p_N$, will show has a solution

$$p_2 = g_2^2(p_3,\ldots,p_N).$$

Since $f_1$ is increasing in $p_1$ and decreasing in $p_2$ it follows that $g_1^1$ is increasing in $p_2$. Moreover, $f_1(g_1^1(p_2,\ldots,p_N),p_2,\ldots,p_N) = \pi_1$ for all $p_2$ and for $j \geq 3$ $f_j(g_1^1(p_2,\ldots,p_N),p_2,\ldots,p_N)$ is, since $g_1^1$ is increasing in $p_2$, decreasing in $p_2$. Thus it follows that (here and below all $f_j$ are evaluted in $(g_1^1(p_2,\ldots,p_N),p_2,\ldots,p_N)$)

$$f_2 = n - \pi_1 - f_3 - \ldots - f_N$$

is increasing in $p_2$ and by the continuity of $g_1^1$ it is also continuous in $p_2$. To see that $\pi_2$ is in the range of this function, we recall the assumption that $\pi_1 \leq \pi_2 \leq \ldots \leq \pi_N$. Letting $p_2 \to 1$ (which eventually means that $p_2 \geq \max(p_3,\ldots,p_N)$) we then eventually have that $f_2 = \max(f_2,\ldots,f_N) \geq (n-\pi_1)/N \geq \pi_2$. On the other hand, letting $p_2 \to 0$ will either imply that $p_2 \leq g_1^1(p_2,\ldots,p_N)$ for some $p_2$ so that $f_2 \leq f_1 = \pi_1 \leq \pi_2$ for this $p_2$ or that $g_1^1$ will also tend to zero. If $n \leq N-2$ the latter would mean that $f_1 \to 0$, a contradiction, and if $n = N-1$ it would imply that $f_j \to 1$, $j \geq 3$, so that $f_1 + f_2 = \pi_1 + f_2 \to 1$. But since $n = N-1$ we have that $\pi_1 + \pi_2 > 1$, so that $f_2 < \pi_2$ for small enough $p_2$ as desired.

Thus, there is a unique pair $p_2 = g_2^2(p_3,\ldots,p_N)$ and

$$p_1 = g_1^2(p_3,\ldots,p_N) = g_1^1(g_2^2(p_3,\ldots,p_N),p_3,\ldots,p_N),$$

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

Now, suppose that the procedure indicated above has been carried out for $l = 2,3,\ldots,k$, $k < N-1$, so that at each step we have found unique values of $p_1,\ldots,p_l$ as continuous functions $g_1^l,\ldots,g_l^l$ of $p_{l+1},\ldots,p_N$ so that $(e1),\ldots,(el)$ are satisfied. It is then straightforward to carry out step $l = k+1$ except for one new difficulty, namely that in order to use the same arguments as for the case $l = 2$, it must be verified that $g_1^k,\ldots,g_k^k$ are increasing as functions of $p_{k+1}$. However, if we assume that increasing $p_{k+1}$ to a larger value $p'_{k+1}$ causes $g_i^k$, $i \in B$, for a nonempty subset, $B$, of $\{1,\ldots,k\}$ to decrease (and the others to increase or stay fixed), we have that the sum $\sum_{i\in B} f_i$ would decrease, a contradiction. To see that this sum must decrease we proceed in two steps. First we increase $p_{k+1}$ and $p_i$ for $i \notin B$, $i \leq k$, to their larger new values and keep all the other $p_i's$ at their original values. Then all $f_i$'s

for $i \in B$ decrease. Second we change the $p_i$'s for $i \in B$ to their smaller new values. This causes all $f_i$'s for $i \in \{1, \ldots, N\} \setminus B$ to increase and thus $\sum_{i \in B} f_i$ to decrease even further as $\sum_{i=1}^{N} f_i$ always remains equal to $n$. It should also be pointed out that in this step of the induction the special treatment of the case $n = N - 1$ above must be extended to the cases $n \geq N - k$; let $p_{k+1}$ tend to zero and assume that $g_j^k$ also tends to zero for $j = 1, \ldots, k$, then, for such an $n$, $f_j, j = k+2, \ldots, f_N$ all tend to one, which implies that $f_{k+1} < \pi_{k+1}$ for small enough $p_{k+1}$ in the same way as for $l = 2$.

Thus we can proceed inductively to eventually get, for each fixed $p_N$, uniquely determined values of $p_1, \ldots, p_{N-1}$ satisfying (e1),...,(eN-1). Since $\sum_{i=1}^{N} f_i = n$, (eN) will be automatically satisfied. The last step is now to calibrate $p_N$ so that $\sum_{i=1}^{N} p_i = n$. However, since all the $p_i$'s, $i = 1, \ldots, N-1$ are increasing functions of $p_N$, it is clear that $\sum_{i=1}^{N} p_i$ is increasing in $p_N$. Letting $p_N \to 0$ implies that this sum also tends to 0 as $\pi_1 \leq \ldots \leq \pi_N$. On the other hand we can at this stage equally well regard $p_2, \ldots, p_N$ as functions of $p_1$ and letting $p_1 \to 1$ implies that $\sum_{i=1}^{N} p_i \to N$ by the same reason. Therefore $p_N$ can be uniquely determined so that this sum equals $n$. Alternatively this can be seen by observing that changing the $p_i$'s in such a way the $\frac{p_i/(1-p_i)}{p_j/(1-p_j)}$ is kept fixed for every $i$ and $j$, the $f_i$'s do not change. □

# 3 A Central Limit Theorem for Conditional Poisson Samples

Let us again consider a Poisson sample

$$s \subseteq \{1, 2, \ldots, N\}$$

with unconditional inclusion probabilities

$$p_i = P(i \in s), \qquad i = 1, \ldots, N.$$

Let $x_i$ be the value for individual $i$, $i = 1, \ldots, N$, of a certain study variable and let

$$t_x = \sum_i x_i$$

be the corresponding population total. In this unconditional case the Horvitz-Thompson estimator of this total is

$$\hat{t}_x = \sum_i \frac{x_i}{p_i} I_i$$

and since all $I_i, i = 1, \ldots, N$ are independent, we may use classical central limit theory for sums of (triangular arrays of) independent random variables to deduce approximate normality of $\hat{t}_x$, under mild conditions. What happens if we condition on the event that $|s| = \sum_i I_i = n$? Can we still argue a normal approximation of $\hat{t}_x$ given that $|s| = n$?

Note that, to get a Horvitz-Thompson estimator, we should use the conditional inclusion probabilities

$$\pi_i = P(i \in s \,|\, |s| = n)$$

instead of $p_i$ and replace $\hat{t}_x$ by

$$\sum_{i \in s} \frac{x_i}{\pi_i},$$

which is an unbiased estimator of $t_x$ under the conditional distribution of $s$. Thus to be as general as possible we let $y_1, \ldots, y_N$ be real numbers and study the conditional distribution of $\sum_i y_i I_i$ given $|s| = n$.

From Cramér-Wold's device (see Billingsley (1968)) it is not hard to deduce that

$$(\sum_i y_i I_i, \sum_i I_i)$$

approximatly has a two-dimensional normal distribution (under suitable conditions). This is naturally formulated as a limit theorem for a sequence of Poisson samples drawn from a sequence of populations with a sequence of $y_i$-values. We shall try to show a corresponding limit theorem for the conditional distribution of $\sum_i I_i y_i$ given that $|s| = \sum_i I_i = n$.

In principle we shall think of a $\nu$-indexed sequence of conditional Poisson samples with conditional inclusion probabilities $\pi_1^{(\nu)}, \ldots, \pi_{N^{(\nu)}}^{(\nu)}$ with sizes $n^{(\nu)} = \sum_i \pi_i^{(\nu)}$ from a sequence of populations of sizes $N^{(\nu)}$ with a sequence of $y_1^{(\nu)}, \ldots, y_{N^{(\nu)}}^{(\nu)}$-values. However, we shall not use any explicit $\nu$-indeces in the formulas in order not to burden the notation. As is shown in Theorem 2.2 it is no restriction to assume that this distribution is realised as the conditional one of a Poisson sample procedure with unconditional inclusion probabilities $p_1, \ldots, p_N$ which sums to $n$ so that the expected sample size in this sample is $\mathbf{E}[|s|] = n$. Thus we assume so in the sequel and to start with, we shall formulate our conditions in terms of these unconditional inclusion probabilities and come back to the conditional ones later.

First we reformulate the problem a second time and study instead the asymptotical distribution of the two-dimensional vector

$$(\sum_i (y_i - \bar{y}_w)(I_i - p_i), \sum_i (I_i - p_i))$$

where $\bar{y}_w$ is the weighted mean

$$\bar{y}_w = \frac{\sum_i y_i p_i (1 - p_i)}{\sum_i p_i (1 - p_i)}.$$

This is convenient because the two components are uncorrelated and the conditional distribution of the first component, given that the second equals 0, coincides with the conditional distribution of a centralised version of the sum we are really interested in.

The question now becomes: When can we deduce that the conditional distribution of $\sum_i (y_i - \bar{y}_w)(I_i - p_i)$ looks like the conditional one in a normal distribution with independent components ie. as the unconditional normal approximation of the same expression with mean 0 and variance

$$\sum_i (y_i - \bar{y}_w)^2 p_i (1 - p_i)?$$

It turns out that this is always the case if, in a triangular formulation,

$$\sum_i p_i (1 - p_i) \to \infty, \tag{2}$$

and

$$\frac{\max_i (y_i - \bar{y}_w)^2 p_i(1-p_i)}{\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i)} \to 0. \tag{3}$$

The main ideas of the proof is to first split

$$\sum_i (y_i - \bar{y}_w)(I_i - p_i) = \sum_i (y_i - \bar{y}_w)^+ (I_i - p_i) - \sum_i (y_i - \bar{y}_w)^- (I_i - p_i)$$

where $x^+ = xI_{\{x>0\}}$ and $x^- = -xI_{\{x<0\}}$, and study the conditional distribution of the vector

$$(\sum_i (y_i - \bar{y}_w)^+ (I_i - p_i), \sum_i (y_i - \bar{y}_w)^- (I_i - p_i)).$$

Both the components of this vector are increasing functions of the inclusion indicators $I_1, \ldots, I_N$ and this can be used in combination with Proposition 6.2 to see that the conditional distributions of the vector given the sample size $|s|$ are stochastically increasing with $|s|$, which in its turn can be utilised to prove a conditional central limit theorem, using Proposition 5.1 below, for appropriately chosen normalised subsequences of the triangular scheme.

However, this also, by the continuous mapping theorem, forces the conditional distribution of the original sum to converge weakly to the right subsequence-independent limit so that Helly's theorem (see Ash (1972), Theorem 8.2.1) can be used to deduce the weak convergence of the sum. We leave the details to Section 7 and give the resulting theorem here:

THEOREM 3.1 *Suppose that to a sequence of conditional Poisson samples correspond sequences of unconditional inclusion probabilities* $\{p_i\}$ *and y-values such that* $\sum_i p_i = n$ *and (2) and (3) are true. Then the conditional distribution of*

$$(\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i))^{-\frac{1}{2}} (\sum_i y_i I_i - \sum_i y_i p_i)$$

*given that* $|s| = n$ *converges weakly to a normal distribution with mean* 0 *and variance* 1. *Furthermore the expectation and variance of this conditional distribution converges to* 0 *and* 1.

# 4 Classical Central Limit Theorems for Poisson Sampling

Here we shall give a multivariate central limit theorem for Poisson samples based on classical theory for sums of i.i.d. random variables. Hajek (1981) used Lindeberg's central limit theorem to show the following:

PROPOSITION 4.1 *If, in a triangular scheme of Poisson samples,*

$$\sum_i p_i(1-p_i) \to \infty$$

*and*

$$\frac{\max_i \ x_i^2 p_i(1-p_i)}{\sum_i x_i^2 p_i(1-p_i)} \to 0$$

*then*

$$(\sum_i x_i^2 p_i(1-p_i))^{-\frac{1}{2}} \sum_i x_i(I_i - p_i)$$

*converges weakly to a standard normal distribution.*

This proposition has the following multivariate extension:

PROPOSITION 4.2 *Consider a triangular scheme of Poisson samples. Suppose that each of the components in a vector version*

$$\sum_i (x_{1i}, \ldots, x_{ki})(I_i - p_i)$$

*is already scaled so that either its variance has limit* 0 *or a finite positive limit and that for the components with positive limits the conditions of Proposition 4.1 above are satisfied. Suppose furthermore that the covariances*

$$\sum_i x_{qi} x_{ri} p_i(1-p_i) \quad , q, r = 1, \ldots, k$$

*converge and denote the limiting covariance matrix by C. Then*

$$\sum_i (x_{1i}, \ldots, x_{ki})(I_i - p_i)$$

*converges to a multivariate normal distribution with expectation* 0 *and covariance matrix C.*

**Remark.** Observe that some (or even all) linear combinations of the components of the limit may be degenerate and have variances 0.

*Proof.* We shall use Cramér-Wold's device. Let $z_i$ be any linear combination of $x_{1i}, \ldots, x_{ki}$. The assumed convergence of the covariance matrices shows that the variance of $\sum_i z_i(I_i - p_i)$ ie.

$$\sum_i z_i^2 p_i(1-p_i)$$

converges. If the limit is 0 the sum $\sum_i z_i(I_i - p_i)$ converges weakly to a degenerate distribution with all mass in 0. If, on the other hand the limit is strictly positive, we may utilise that

$$\max_i x_{ji}^2 \; p_i(1-p_i) \to 0, \qquad j = 1, \ldots, k$$

and, using e.g. Cauchy-Schwarz' inequality, bound $\max_i z_i^2 p_i(1-p_i)$ , so that it can be seen to converge to 0. But this means that the required one dimensional asymptotic normality of $\sum_i z_i(I_i - p_i)$ follows from Proposition 4.1. □

# 5 A Conditional Central Limit Theorem

We are going to use the results in Nerman (1997) which reformulated for our purposes implies the following

PROPOSITION 5.1 *Consider a sequence of random vectors $(X_n, Y_n) \in R^{k+r}$ which converges weakly to a multivariate normally distributed $(X, Y)$-vector, such that $Y$ has an r-dimensional Lebesgue density. Assume also that $Y_n$ are discrete with positive $P(Y_n = y)$ exactly for $y \in D_n$, and that the conditional distributions of $X_n$ given $Y_n = y$ are stochastically monotone in $y \in D_n$ for each $n$. Then, if $y_n \in D_n$ converges to $y_0$, the conditional distribution of $X_n$ given $Y_n = y_n$ converges weakly to the conditional normal distribution of $X$ given $Y = y_0$. (It is understood that the versions of the conditional distributions of $X$ given $Y = y$ are chosen as the natural ones so that weak continuity at $y_0$ is ensured.)*

*If, in addition, the variances of the components of $X_n$ converge to those of $X$, then the conditional expectations, variances and covariances of $X_n$ given $Y_n = y_n$ converge to the those of $X$ given $Y = y_0$.*

# 6 The Conditional Distributions Are Stochastically Increasing

We need to prove the fact that the conditional distributions of the vector $(I_1, \ldots, I_N)$ of inclusion indicators given $|s| = n$ are stochastically increasing in $n$. (Here, and for the rest of this section we refrain from the condition that $\sum_{i=1}^N p_i p_i = n$. On the contrary, the $p_i$'s are assumed to be fixed throughout the section.) This is intuitively obvious but appears to be algebraically messy to prove. Therefore we are going to use a Markov chain idea, which will also suggest a practical method of actually choosing a conditional Poisson sample. The latter implication will be treated in Section 9.

Consider a discrete time Markov chain, $\{u_t\}_{t=0}^{\infty}$, on the state space, $\mathcal{A}_n$, the set of all samples of size $n$. Let the chain start in any state, $u_0$, and let the transitions be given by the following procedure:

At each time, $t$, do the following.

(i) Choose one of the individuals in $u_t$ at random.

(ii) Choose one individual, from the whole population, according to the distribution $\{\theta_i/c\}_{i=1}^N$, where $\theta_i = p_i/(1-p_i)$ and $c = \sum_{i=1}^N \theta_i$. Note that this individual might be the same as the one chosen at (i).

(iii) If the individual chosen at (ii) is not already in $u_t$, then exchange the two individuals.

PROPOSITION 6.1 *The conditional distribution of a Poisson sample, $s$, given $|s| = n$, i.e. the distribution, $\mu$, given by*

$$\mu(s_0) = \frac{1}{P(\mathcal{A}_n)} \prod_{i \in s_0} p_i \prod_{j \notin s_0} (1 - p_j)$$

*is the unique stationary distribution of the Markov chain specified above.*

*Proof.* Let $p(s_0, s_1)$, $s_0, s_1 \in \mathcal{A}_n$, denote the transition probabilities for $\{u_t\}$. If $s_0$ and $s_1$ differ by exactly one individual, say $i \in s_0 \setminus s_1$ and $j \in s_1 \setminus s_0$, then

$$p(s_0, s_1) = \frac{1}{n}\frac{\theta_j}{c}$$

and

$$p(s_1, s_0) = \frac{1}{n}\frac{\theta_i}{c}$$

so that

$$\mu(s_0)p(s_0, s_1) = \frac{1}{ncP(\mathcal{A}_n)} \prod_{k \in s_0 \cup s_1} p_k \prod_{l \notin s_0 \cup s_1} (1 - p_l) = \mu(s_1)p(s_1, s_0).$$

But since $p(s_0, s_1) = 0$ if $s_0$ and $s_1$ differ by more than one individual, the relation

$$\mu(s_0)p(s_0, s_1) = \mu(s_1)p(s_1, s_0)$$

holds for every $s_0$ and $s_1$. Hence $\mu$ is a stationary distribution for $\{u_t\}$. Uniqueness follows from the obvious irreducibility of the process. □

Using Proposition 6.1 and a simple coupling argument, the desired stochastic monotonicity will follow.

PROPOSITION 6.2 *The conditional distributions of* $(I_1, \ldots, I_N)$ *given* $|s| = \sum_i I_i = n$ *are stochastically (strictly) increasing in* $n$.

*Proof.* We are going to construct two Markov chains, $\{u_t^m\}$ and $\{u_t^{m+1}\}$ both having transitions described by (i),(ii) and (iii) for $n = m$ and $n = m+1$ respectively. These are going to be coupled in such a way that $u_t^m \subset u_t^{m+1}$ for every $t$ and so that the pair $(u_t^m, u_t^{m+1})$ is also a Markov chain. Since, with obvious notation, this will mean that

$$I_1^m(t) \leq I_1^{m+1}(t), \ldots, I_N^m(t) \leq I_N^{m+1}(t)$$

for every $t$, it will also imply that the stationary distribution of $\{u_t^m\}$ is stochastically smaller than the stationary distribution of $\{u_t^{m+1}\}$ as desired. Also, since in the stationary distribution of $(u_t^m, u_t^{m+1})$ any pair, $(s, s')$, of samples of sizes $m$ and $m+1$ such that $s \subseteq s'$ has positive probabilty, the inequality is strict.

Now let us do the construction of the two processes. Let $u_0^m \subset u_0^{m+1}$. Let the transitions of $\{u_t^{m+1}\}$ be as described by (i), (ii) and (iii) for $n = m + 1$. Couple the transitions of $\{u_t^m\}$ to the ones of $\{u_t^{m+1}\}$ in the following way. At time $t = 1$, choose the same individual at (i) for the two processes if this individual is in $u_0^m$. If not, choose an individual in $u_0^m$ at random. At (ii), choose the same individual for the $\{u_t^m\}$-process as for the $\{u_t^{m+1}\}$-process. Now do (iii) as usual. It is clear that $u_1^m \subset u_1^{m+1}$ and by continuing in the same way for $t = 2, 3, \ldots$, we have that $u_t^m \subset u_t^{m+1}$ for every $t$. It is easily checked that $\{u_t^m\}$ has the behavior described by (i), (ii) and (iii) for $n = m$. The proof is complete. □

**Note**. The technique used in this section is used in Liggett (1985), Section VIII.2 on the exclusion process, with the difference that this reference uses a continuous time setting.

# 7 Proof of Theorem 3.1

It is no restriction to assume the $y$-values to be scaled so that the

$$\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i) = \sum_i p_i(1-p_i)$$

so that we may multiply by

$$\gamma = (\sum_i p_i(1-p_i))^{-\frac{1}{2}}$$

in order to to standardise both components in

$$(\sum_i (y_i - \bar{y}_w)(I_i - p_i), \sum_i (I_i - p_i)).$$

Think of our underlying triangular scheme as indexed by $\nu$. Consider any subsequence of $\{\nu\}$ and use compactness to substract from this a further subsequence $\{\nu'\}$ so that the variances of the components of

$$\gamma(\sum_i (y_i - \bar{y}_w)^+ (I_i - p_i), \sum_i (y_i - \bar{y}_w)^- (I_i - p_i), \sum_i (I_i - p_i))$$

converge to $\sigma^2, 1-\sigma^2$ and 1 and so that the covariance between the first and the third component of the vector converges to $\kappa$. Then the covariance beetween the second and the third must also converge to $\kappa$. The first two components are automatically uncorrelated. Denote the corresponding limiting subsequence dependent covariance matrix by $C$.

Now, Proposition 4.2 implies that

$$\gamma(\sum_i (y_i - \bar{y}_w)^+ (I_i - p_i), \sum_i (y_i - \bar{y}_w)^- (I_i - p_i), \sum_i (I_i - p_i))$$

converges weakly, as $\nu' \to \infty$, to a normal distribution with mean 0 and covariance matrix $C$.

Using the well known fact that if $X$ is stochastically smaller than $Y$ and $f$ is an increasing function, then $f(X)$ is stochastically smaller than $f(Y)$ it follows from Proposition 6.2 that the conditional distributions of the first two components given $|s| = k$ are stochastically increasing in $k$. Thus Proposition 5.1 shows that the conditional distribution of

$$\gamma(\sum (y_i - \bar{y}_w)^+ (I_i - p_i), \sum (y_i - \bar{y}_w)^- (I_i - p_i))$$

given that $\sum_i (I_i - p_i) = 0$ converges weakly, as $\nu' \to \infty$, to the normal distribution recieved from conditioning this three-dimensional normal distribution with its third component being 0. Observe also that since the unconditional variances converge to those of the normal limit, the last part of Proposition 5.1 applies and the conditional expectations, variances and covariance converge as well. Thus, from the continuous mapping theorem of weak convergence it follows that also the conditional distribution of the difference of the first two components

$$\gamma(\sum_i (y_i - \bar{y}_w)^+ (I_i - p_i) - \sum_i (y_i - \bar{y}_w)^- (I_i - p_i) = \gamma \sum_i y_i (I_i - p_i)$$

converges weakly, as $\nu' \to \infty$, to the corresponding conditional difference distribution in the three-dimensional normal distribution above. It also follows that the conditional expectation and variance converge to the ones of the limit. However, this limiting distribution is always a normal distribution with mean 0 and variance 1. Since this limit does not depend on which original subsequence of $\{\nu\}$ we substracted, we can apply Helly's Theorem to finish the proof of Theorem 3.1.

# 8 The Conditions in Terms of $\{\pi_i\}$

The centering constant in Theorem 3.1 is a function of the unconditional inclusion probabilities $p_i$. The weights $p_i(1-p_i)$ in the weighted average $\bar{y}_w$ and in the normalising sum in Theorem 3.1 are derived from the unconditional inclusion probabilities. Moreover, $p_i(1-p_i)$ are also used in conditions (2) and (3). It is natural to ask whether in all these expressions the unconditional inclusion probabilities may be substituted by the conditional ones, $\pi_i$? The answer is yes, as we shall demonstrate below; we can substitute all $p_i$ by $\pi_i$ in (2), in (3), in the definition of $\bar{y}_w$, in the centering or in the normalising constant of Theorem 3.1 and still get a valid theorem.

First we observe that since Theorem 3.1 implies that the conditional mean of the converging expression tends to 0, we get that

$$(\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i))^{-\frac{1}{2}} (\sum_i y_i \pi_i - \sum_i y_i p_i) \to 0.$$

This of course means that we can substitute the centering constant in Theorem 3.1 by the conditional mean $\sum_i y_i \pi_i$ of $\sum_i y_i I_i$.

Next, let us turn to the relation between $\{p_i\}$ and $\{\pi_i\}$. It is natural to believe that if

$$\sum_i p_i(1-p_i) \to \infty$$

then $\pi_i/p_i$ and $(1-\pi_i)/(1-p_i)$ both should converge to 1 for all $i$. Below we shall show that this is indeed true and that these convergences are uniform, beginning with the former.

Define $\pi_i(k)$ to be the conditional inclusion probability of individual $i$ in the Poisson sample given that the sample size $|s| = k$, $i, k = 0, \ldots, N$. We know from Proposition 6.2 that $\pi_i(k)$ increases in $k$ for any $i$. Observe also that

$$\pi_i(k) = \frac{P(\sum_{j;j\neq i} I_j = k-1)}{P(\sum_j I_j = k)} p_i \tag{4}$$

These two facts imply the inequalities

$$\frac{P(\sum_{j;j\neq i} I_j \leq (n-1))}{P(\sum_j I_j \leq n)} \leq \frac{\pi_i}{p_i} \leq \frac{P(\sum_{j;j\neq i} I_j \geq (n-1))}{P(\sum_j I_j \geq n)}. \tag{5}$$

The left inequality follows from (4) and the inequality of the weighted average

$$\frac{\sum_{k\leq n} \pi_i(k) P(\sum_j I_j = k)}{\sum_{k\leq n} P(\sum_j I_j = k)} \leq \pi_i(n) = \pi_i$$

which in its turn is a consequence of the monotonicity of $\{\pi_i(k)\}$. The right follows from a similar weighted average of $\pi_i(k)$ for $k \geq n$.

Observe that, for all $i$,

$$E[(I_i - p_i)^3] \leq E[(I_i - p_i)^2]$$

so that we can use Berry's inequality for sums of independent, but not necessarily equi-distributed random variables, Feller (1966), page 521 to approximate the

nominators and denominators in (5). Recall that $\gamma$ denotes $(\sum_i p_i(1-p_i))^{-\frac{1}{2}}$. It is straightforward to see that the absolute differences

$$|P(\sum_j I_j \leq n) - \frac{1}{2}|$$

$$|P(\sum_{j;j\neq i} I_j \leq n-1) - \frac{1}{2}|$$

$$|P(\sum_j I_j \geq n) - \frac{1}{2}|$$

and

$$|P(\sum_{j;j\neq i} I_j \geq n-1) - \frac{1}{2}|$$

are all bounded by $C'\gamma$, where $C'$ is a universal constant. It follows from this that, for some other constant $C''$ (and $\sum_i p_i(1-p_i)$ large enough),

$$|\frac{\pi_i}{p_i} - 1| \leq C''\gamma.$$

By a symmetric argument, we also get that

$$|\frac{(1-\pi_i)}{(1-p_i)} - 1| \leq C''\gamma.$$

Thus there is a third universal constant $C$ such that

$$|\frac{\pi_i(1-\pi_i)}{p_i(1-p_i)} - 1| \leq C\gamma. \tag{6}$$

This in its turn implies that

$$|\frac{\sum_i \pi_i(1-\pi_i)}{\sum_i p_i(1-p_i)} - 1| \leq C\gamma$$

so that if $\sum_i p_i(1-p_i)$ or $\sum_i \pi_i(1-\pi_i)$ converges to infinity so does the other, and

$$\frac{\sum_i \pi_i(1-\pi_i)}{\sum_i p_i(1-p_i)} \to 1.$$

Thus (2) is equivalent to

$$\sum \pi_i(1-\pi_i) \to \infty.$$

Similarly it follows from (6) that, for any $a$ $(= a^{(\nu)})$,

$$\frac{\sum_i (y_i-a)^2 p_i(1-p_i)}{\sum_i (y_i-a)^2 \pi_i(1-\pi_i)} \to 1, \tag{7}$$

provided that either of the sums $\sum_i p_i(1-p_i)$ or $\sum_i \pi_i(1-\pi_i)$ converges to infinity.

Now, suppose that the average $\bar{y}_\pi$ is defined from the weights $\pi_i(1-\pi_i)$ instead of the original weights $\{p_i(1-p_i)\}$ used in the definition of $\bar{y}_w$. Then the facts that $\bar{y}_\pi$ minimizes the expression

$$\sum_i (y_i-a)^2 \pi_i(1-\pi_i)$$

as a function of $a$, and that $\bar{y}_w$ minimizes

$$\sum_i (y_i - a)^2 p_i(1-p_i),$$

show the two inequalities

$$\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i) \leq \sum_i (y_i - \bar{y}_\pi)^2 p_i(1-p_i),$$

and

$$\sum_i (y_i - \bar{y}_\pi)^2 \pi_i(1-\pi_i) \leq \sum_i (y_i - \bar{y}_w)^2 \pi_i(1-\pi_i).$$

But (2), in the light of (7), implies that the left side of the first inequality is asymptotically equivalent to the right of the second and vice versa the right side of the first inequality is asymptotically equivalent to the left of the second. Thus (2) clearly forces all four expressions to be asymptotically equivalent, so that the square root of anyone of them can be used to normalise in Theorem 3.1. Let us also remark that a fifth alternative standardisation of $\sum_i y_i I_i$ follows from the second moment convergence of the same theorem. We can use the standard deviation of the conditional distribution of $\sum_i y_i I_i$. Written in Yates-Grundy's form (see Särndal *et al.*) using bivariate conditional inclusion probabilities $\pi_{ij}$, this equals

$$(\sum_{ij} (\pi_{ij} - \pi_i \pi_j)(y_i - y_j)^2)^{\frac{1}{2}} \tag{8}$$

Next, we shall argue that, given (2), (3) is equivalent to

$$\frac{\max_i \ (y_i - \bar{y}_\pi)^2 p_i(1-p_i)}{\sum_i (y_i - \bar{y}_\pi)^2 p_i(1-p_i)} \to 0. \tag{9}$$

To see this, observe that

$$\sum_i (y_i - \bar{y}_\pi)^2 p_i(1-p_i) = \sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i) + (\bar{y}_w - \bar{y}_\pi)^2 \sum_i p_i(1-p_i),$$

so that we get, from the asymptotic equivalence of the left side and the first sum of the right, that

$$\frac{\max_i \ (\bar{y}_w - \bar{y}_\pi)^2 p_i(1-p_i)}{\sum_i (y_i - \bar{y}_w)^2 p_i(1-p_i)} \to 0,$$

which in its turn (using the inequality $(x+y)^2 \leq 2(x^2+y^2)$) yields that (2) and (3) imply (9). A symmetric argument shows that (2) and (9) imply (3) so that (3) and (9) are exchangeble in Theorem 3.1. Of course, using (7), we may also substitute all $p_i(1-p_i)$ in (3) as well as in (9) by $\pi_i(1-\pi_i)$.

Finally, we return to the Horvitz-Thompson estimation. Are we allowed to use $\sum_i \frac{x_i}{\pi_i} I_i$ instead of $\sum_i \frac{x_i}{p_i} I_i$ or vice versa? That is, assume that the assumptions of Theorem 3.1 are satisfied so that the conditional distribution of one of them is asymptotically normal. Are the variances of the two estimators asymptotically equivalent and can normality be deduced for the other estimator too? These are quite tricky questions, but by letting $x_i = p_i$, for all $i$, or $x_i = \pi_i$, for all $i$, and using only two different values for the $p_i$'s, we may end up in a situation where one of the estimators has an approximate normal distribution whereas the other one has a degenerate distribution. Thus such deductions are not generally possible.

# 9 Implementation of the Sampling Procedure

When using the conditional Poisson sampling procedure described in this text in practice, two computational problems occur.

(a) Calculating the $p_i$'s.

(b) Choosing the sample.

As we saw in the previous section we can, for large samples, let $p_i$ equal the desired conditional inclusion probability. This does not give the exact solution to (1) but it gives a good approximation. For small sample sizes it should be possible to make the exact calculation of the $p_i$'s, but already for moderate sample sizes and different inclusion probabilities this seems to be a tough task. To come up with an approximation procedure for this case seems like a good problem for an interested student. Using a good computer and a suitable recursion it should be a bit more straightforward to find numerical values for the $\pi_i$'s given the $p_i$'s.

Now consider (b). Using central limit heuristics (i.e. a not proven local central limit theorem) it follows that doing this step by the immediate method of choosing Poisson samples and rejecting them until exactly $n$ individuals are chosen, would on average require the order of $N\sqrt{\sum_i p_i(1-p_i)}$ operations for a computer. In reasonable situations the $p_i$'s are bounded away from 1 so that $\sum_i p_i(1-p_i)$ is at most of the order $n$. Thus it should take on average the order of $N\sqrt{n}$ operations to choose the sample. If we e.g. consider sampling from a population of the order of $10^4$ individuals, then $N\sqrt{n}$ is at most of the order $10^6$ and modern computers will have no problem to fulfil the task fairly quickly. Similarly, in larger populations with smaller sample sizes there will be no problem. However, for large samples from very large populations it is worth to consider other procedures. One natural attempt in this way is to use the method suggested to us in Section 6, i.e. actually use the Markov chain introduced there. Then the question of how long time it takes to come close to stationarity arises. To answer this question we will again use a coupling technique. Let $\{u_t\}$ be the process we are actually interested in, i.e. a Markov chain starting in some fixed state, $u_0$, and with transitions governed by (i), (ii) and (iii) of Section 6. Let $\{u'_t\}$ be another Markov chain starting in stationarity, i.e. such that $u_0$ has the distribution $\mu$ of Proposition 6.1. The transitions of $\{u'_t\}$ will also satisfy (i), (ii) and (iii) but are to be coupled with those of $\{u_t\}$ in the following way at time $t = 1, 2, \ldots$. If the individual chosen in $u_t$ at (i) is also in $u'_t$, then choose the same individual in $u'_t$. If not, choose one of the individuals in $u'_t \setminus u_t$ at random. At (ii), choose the same individual for the $\{u'_t\}$-process as for the $\{u_t\}$-process and then do (iii). Let $T$ be the *coupling time*, i.e. the first time the two processes meet. Formally we have

$$T = \inf\{t : u_t = u'_t\}.$$

It is natural to measure the distance between the distribution, $\mu^{(t)}$, of $u_t$, and $\mu$ by the *total variation norm*:

$$||\mu^{(t)} - \mu|| = \frac{1}{2} \sum_{s_0 \in \mathcal{A}_n} |\mu^{(t)}(s_0) - \mu(s_0)|.$$

Using the well known coupling inequality (cf. for instance Lindvall (1992), (2.8), page 12) gives the following bound on the total variation norm.

$$||\mu^{(t)} - \mu|| \leq Pr(T > t).$$

Assuming that the two processes start in the "worst possible" way with no individuals in common, we may rewrite $T$ as

$$T = \sum_{k=1}^{n} T_k$$

where $T_k$ is defined as $\tau_k - \tau_{k-1}$, where $\tau_k = \inf\{t : |u_t \cap u'_t| = k\}$, the first time $u_t$ and $u'_t$ have $k$ individuals in common. Observe that $|u_t \cap u'_t|$ never decreases and that if $u_t$ and $u'_t$ have $k$ individuals in common at a certain time, $t$, then the probability that $u_{t+1}$ and $u'_{t+1}$ will have $k+1$ individuals in common is at least

$$\frac{n-k}{n} K_n^{-1}$$

where

$$K_n = (1 - \frac{1}{c} \sup_{s_0 \in \mathcal{A}_n} \sum_{i \in s_0} \theta_i)^{-1}$$

(remember that $\theta_i = p_i/(1-p_i)$ and $c = \sum_{i=1}^{N} \theta_i$) for this will be the case if the individuals chosen at (i) are not identical and the individual chosen at (ii) is not in $u_t \cap u'_t$. Thus, using the Markov property of $(u_t, u'_t)$, it follows that $T$ is stochastically not larger than the sum of $n$ independent geometric random variables whose expectations are at most

$$K_n \frac{n}{n-k}$$

for $k = 1, 2, \ldots, n$ so that we have

$$\mathbf{E}[T] \leq K_n \sum_{k=1}^{n} \frac{n}{n-k}$$

which is of the order $K_n n \log n$ if $n$ is large. By calculating the second moments of the $T_k$'s and using Chebyshev's inequality one can prove that

$$Pr(T > (1+\epsilon) K_n n \log n) \to 0$$

as $n \to \infty$ for any $\epsilon > 0$. The details of the last argument are carried out in Carlsson (1996), Theorem 3.2 and are therefore omitted here. The coupling inequality now implies that

$$||\mu^{((1+\epsilon)K_n n \log n)} - \mu|| \to 0$$

as $n \to \infty$. One says that $K_n n \log n$ is an *upper bound* for the convergence rate of the distribution of $u_t$.

In reasonable situations the $K_n$'s are bounded (and usually not much larger than 1) so that the upper bound is of the order $n \log n$. Now, since a computer would need $\log_2 N$ operations to classify the choice at (ii), it would take the order of $n \log N \log n$ operations to carry out $n \log n$ transitions of the Markov chain. If $n$ is large and $N$ is much larger than $n$, this is a considerable gain compared to the "naive" method. One drawback is of course that the resulting sample will not have exactly the distribution $\mu$, but if we do for instance $2\mathbf{E}[T]$ transitions and $n$ is large, it will be extremely close. Note also that we can of course do much better than starting the process in a fixed state. Letting $u_0$ have a distribution which is at least

fairly close to $\mu$ will improve the result considerably. For instance we could use a sample drawn with the technique used in Rosén (1995) or Olsson (1995) for $u_0$.

**Acknowledgment.** We are grateful to Olle Häggström for supplying the Markov chain idea used to prove Proposition 6.2 and for valuable comments on the manuscript. We are also grateful to the referee for valuable comments and for pointing out a gap in the proof of Theorem 2.2.

## References

Ash, R. B. (1972). *Real Analysis and Probability,* Academic Press, San Diego.

Billingsley P. (1968). *Convergence of Probability Measures,* Wiley, New York.

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

Carlsson, H. (1996). *On Shuffling Decks with Black and White Cards,* Phd Thesis, Dept. of Mathematics, University of Göteborg and Chalmers University of Technology.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications,* Wiley, New York.

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

Hajek, J. (1981). *Sampling From a Finite Population,* Marcel Dekker, New York.

Horvitz, D. G. & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe, *J. Amer. Statist. Assoc.* **47**, 663-685.

Liggett, T. M. (1985). *Interacting Particle Systems,* Springer, New York.

Lindvall, T. (1992). *Lectures on the Coupling Method,* Wiley, New York.

Nerman, O. (1997). Stochastic Monotonicity and Conditioning in the Limit, *Scand. J. Statist.*, to appear.

Olssson, E. (1995). *Sequential Poisson Sampling,* Research Report 182, Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University.

Rosén, B. (1995). *Asymptotic Theory for Order Sampling,* Statistics Sweden R & D Report 1995:1.

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