Journal of Official Statistics, Vol. 14, No. 2, 1998, pp. 149–162

Sequential Poisson Sampling

Esbjörn Ohlsson[1]

Poisson sampling is a simple way to draw a probability proportional to size (pps) sample from a finite population. It also offers an easy way to update a sample while retaining as many units as possible from the previous sample, and/or to minimize overlap of different samples. A drawback of Poisson sampling is the random sample size. We present a fixed size alteration of Poisson sampling, *sequential Poisson sampling*, designed for, and used in, the Swedish Consumer Price Index (CPI). We show that the respective estimators associated with ordinary and sequential Poisson sampling, are both asymptotically normally distributed and unbiased as well as equally efficient. Simulations on CPI data verify approximate unbiasedness and approximate equality of variances, plus equally good performance of associated estimators of variance. Therefore, sequential Poisson sampling is preferable, because of the fixed size.

Key words: Probability proportional to size; sample coordination; sample updating; overlap control; asymptotic normality; consumer price index.

1. Introduction

Consider a finite population $U = \{1, 2, \ldots, N\}$ recorded in a list frame, together with some positive auxiliary variable $\mathbf{p} = (p_1, p_2, \ldots, p_N)$. In a typical application we have a stratified design, in which case what we consider here is a single, arbitrary stratum. We assume that $p_i > 0$ for all i and that (within the stratum)

$$\sum_1^N p_i = 1 \tag{1.1}$$

Within strata, we want to sample units with probabilities proportional to p_i. We shall think of p_i as some measure of the size of unit i, and say that a sampling procedure is strictly probabilities proportional to size (pps) if

$$\Pr(i \in s) = np_i, \qquad i = 1, 2, \ldots, N \tag{1.2}$$

where $i \in s$ denotes that unit i is included in the sample s, and n is the desired sample size. Sigman and Monsour (1995) note the use of pps sampling for business surveys, especially for price index estimation. In this kind of application, n is typically of moderate or large size, cf., Dalén and Ohlsson (1995). This is in contrast to the case with pps sampling in multi-stage surveys, where $n = 1$ or 2 is common.

[1] Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden.
e-mail: esbj@matematik.su.se. Phone: +46 8 16 45 58. Fax: +46 8 612 67 17. http://www.matematik.su.se/matstat/

Acknowledgments: The author is grateful to Jörgen Dalén and Bengt Rosén, Statistics Sweden, for valuable discussions during the course of this work.

In strata with moderate size N, there may be a substantial loss in efficiency if sampling is done with replacement, see Section 4 for some examples. Hence, there is a need for techniques for pps sampling without replacement.

A simple without replacement pps procedure is *Poisson sampling*. To each unit in the frame we associate an independent random number, denoted by X_i for unit i. Each X_i is uniformly distributed on the interval [0,1]. Unit i is included in the sample if

$$X_i \leq np_i \tag{1.3}$$

Poisson sampling is obviously strictly pps. Examples of the use of Poisson sampling include the U.S. Bureau of the Census's Annual Survey of Manufacturers (Ogus and Clark 1971) and the Swedish CPI before 1989 (Ohlsson 1990).

In a repeated survey we often want a large overlap of subsequent samples, for efficiency and cost reduction. On the other hand, we want the sample to reflect changes in the population such as births, deaths and changes in size measure or classification. Brewer, Early, and Joyce (1972) suggested the use of Poisson sampling in connection with *permanent random numbers* (PRN) to solve this problem. The idea is to let the X_i from the first sample be permanently associated with the population units. The next Poisson sample is drawn from the updated population using the same random numbers as before. Hence, in (1.3) n and p_i will vary from time to time to reflect population and design changes, while a sample overlap is obtained because X_i will be the same on all sampling occasions. The amount of overlap will, of course, depend on the amount of changes in the population. We may also change stratification and still retain a large part of the old sample, due to the use of PRN.

With PRN it is also possible to reduce the overlap between samples for different surveys taken from the same frame, even if the surveys have different design. This type of overlap control is an important tool for spreading respondent burden in business surveys. Poisson sampling with PRN is used for this purpose in New Zealand, see Templeton (1990). A PRN technique for srswor (simple random sampling without replacement) was suggested by Atmer, Thulin, and Bäcklund (1975) and is now used for most business surveys at Statistics Sweden. For an overview of sample coordination with PRN in different countries, see Ohlsson (1995a). A drawback of Poisson sampling is that the realized sample size m is random, with expectation n. Since m is approximately Poisson distributed, with variance n, the deviations from the desired size n may be considerable. With moderate sample sizes within a large number of strata the result may be serious deviations from an optimal allocation. We may also have to increase the sample size in some strata in order to avoid the possibility of getting empty samples.

We conclude that it is desirable to have a modification of Poisson sampling that yields a fixed sample size n, while still allowing the use of PRN for sample updating and overlap control. Preferably, the method should be as simple to use as Poisson sampling. Such a procedure, *sequential Poisson sampling*, is presented in Section 2.2 below. An additional advantage of sequential Poisson sampling is the possibility to get a fixed number of in-scope units in the sample even if the frame contains out-of-scope units, see Section 2.4. From 1989 sequential Poisson sampling has replaced Poisson sampling as the major sampling procedure for the Swedish Consumer Price Index (CPI), see Ohlsson (1990).

Brewer et al. (1972) gave a technique called *collocated sampling* for reducing, but not

eliminating, the variability of the Poisson sample size. As shown in Ohlsson (1995a), this technique is not as well suited for PRN sample coordination as ordinary and sequential Poisson sampling.

In the literature there is an abundance of fixed size pps procedures, see Brewer and Hanif (1983) for an overview. However, algorithms for updating samples drawn with these procedures are not available for general n. Sunter (1989) gives a procedure which in principle allows general n, but solves only the restricted problem of updating size measures, keeping strata and allocation fixed. For the special case $n = 1$, techniques for updating pps samples are given by Keyfitz (1951) and Kish and Scott (1971).

A problem with sequential Poisson sampling is that no closed expressions can be given for the first and second order inclusion probabilities. Hence, the standard theory for unbiased (Horvitz-Thompson) estimators cannot be used. Nevertheless, as we shall see, estimation and variance estimation is as simple for sequential as for ordinary Poisson sampling.

Though sequential Poisson sampling is very simple to use in practice, its theory is quite intricate. A main purpose of this article is to provide the theory for inference from sequential Poisson samples, pertaining to point estimation, variance and interval estimation (using approximate normality). This is done by giving asymptotic theory and (in case of point and variance estimation) simulation results. We also compare the efficiency of sequential and ordinary Poisson sampling. In particular, the estimators for both procedures are shown to be asymptotically normal, asymptotically unbiased and asymptotically equally efficient.

In Section 2 we present sequential Poisson sampling, estimators and variance estimators. Stringent asymptotic results are presented in Section 3, while proofs are postponed to the appendix. Section 4 contains a report on simulation results, followed by our conclusions in Section 5.

A Historical Note. Ogus and Clark (1971) report that Poisson sampling has been used at the U.S. Bureau of the Census since (at least) 1959. The earliest appearances of the name ''Poisson sampling'' in the literature seem to be in articles by Hájek (1960) for equal probabilities and Hájek (1964) for varying probabilities. The name probably goes back to the term ''Poisson trials,'' mentioned by Feller (1950, p. 234). Särndal, Swensson, and Wretman (1992) introduced the name ''Bernoulli sampling'' for equal probability Poisson sampling.

2. Ordinary and Sequential Poisson Sampling

In the sequel, we denote (ordinary) Poisson sampling by PS and sequential Poisson sampling by SPS. In order for (1.2) to be possible we assume from now on that

$$np_i \leq 1, \qquad i = 1, 2, \ldots, N \tag{2.1}$$

In practice this can always be achieved by transferring exceedingly large units to a ''take-all'' stratum. The object of the survey is to estimate the total of the study variable $\mathbf{y} = (y_1, y_2, \ldots, y_N)$, i.e.,

$$Y = \sum_1^N y_i \tag{2.2}$$

2.1. Poisson sampling

PS was described in the preceding section. Note that the probability of getting an empty sample is

$$\Pr(m = 0) = \prod_{i=1}^{N}(1 - np_i) \leq e^{-n} \tag{2.3}$$

It cannot be recommended to use PS unless this probability is negligible, which we will assume from now on. The unbiased (Horvitz-Thompson) estimator of Y is

$$\hat{Y}_{HT} = \frac{1}{n}\sum_{i \in s}\frac{y_i}{p_i} \tag{2.4}$$

The following expression for the variance of $\hat{Y}_{HT}$ is easy to derive, see e.g., Särndal et al. (1992, p. 86),

$$Var(\hat{Y}_{HT}) = \frac{1}{n}\sum_{i=1}^{N}(1 - np_i)\left(\frac{y_i}{p_i}\right)^2 p_i \tag{2.5}$$

As suggested by (2.5), $\hat{Y}_{HT}$ is usually of poor precision. Brewer et al. (1972), suggested a natural alternative estimator

$$\hat{Y}_R = \begin{cases} \dfrac{1}{m}\displaystyle\sum_{i \in s}\frac{y_i}{p_i} & \text{if } m > 0 \\ 0 & \text{if } m = 0 \end{cases} \tag{2.6}$$

The conventional assignment of the value 0 to $\hat{Y}_R$ in case $m = 0$ is necessary to make the estimator well defined. Note that $\hat{Y}_R$ is the ordinary ratio estimator using $\mathbf{p}$ as auxiliary information. Hence, Result 7.3.1 in Särndal et al. (1992, p. 248) can be used to derive the following approximate expression for the variance of $\hat{Y}_R$,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{N}(1 - np_i)\left(\frac{y_i}{p_i} - Y\right)^2 p_i \tag{2.7}$$

Formula (2.7) can also be obtained from (4.2.26) in Brewer and Hanif (1983), by replacing $\Pr(m = 0)$ by 0 there. To get confidence intervals we must further know that $\hat{Y}_R$ is approximately normally distributed.

Proposition 2.1. The Poisson sampling estimator $\hat{Y}_R$ is, under general conditions, approximately normally distributed with mean Y and variance σ^2.

This somewhat loosely formulated proposition is justified by a heuristic proof in the appendix, by a strict asymptotic result, Theorem 3.1, and by simulations in Section 4. A slight improvement of σ^2 is given in Remark 2.1.

If the ''fpc'' $(1 - np_i)$ is neglected, a comparison of (2.5) and (2.7) shows that $\hat{Y}_R$ reduces the variance by Y^2/n. The simulation studies in Sunter (1977) and our Section 4 indicate that $\hat{Y}_R$ is indeed of much better precision than $\hat{Y}_{HT}$ in situations where pps sampling is appropriate. Särndal (1996) also advocates the use of $\hat{Y}_R$ and the approximation (2.7), which correspond to his Equations (31) and (32).

2.2. *Sequential Poisson sampling*

The disadvantages of the random Poisson sample size were discussed in the introduction. We now present SPS, our fixed size alteration of PS. From the random numbers X_i we form the *transformed random numbers*

$$\xi_i = X_i/p_i \tag{2.8}$$

The PS inclusion rule (1.3) is trivially equivalent to: Include unit i in the sample if (and only if) $\xi_i \leq n$. This formulation of PS suggests the alteration to select the n units having the smallest ξ_i.

Definition 2.1. A sample is said to be drawn by *sequential Poisson sampling* (SPS) of size n if it consists of the n units with the smallest transformed random numbers ξ_i, where ξ_i is defined in (2.8).

SPS was introduced as an outlet sampling procedure in the Swedish Consumer Price Index in 1989, see Ohlsson (1990). It is easy to show by example that, unfortunately, SPS is not *strict* pps, again see Ohlsson (1990). From its close relation to Poisson sampling it is natural to conjecture that SPS is *approximately* pps, though. The simulation results in Section 4 give strong support for this conjecture. This leads us to consider the following estimator in connection with SPS,

$$\hat{Y}_S = \frac{1}{n}\sum_{i\in s}\frac{y_i}{p_i} \tag{2.9}$$

Proposition 2.2. The sequential Poisson sampling estimator $\hat{Y}_S$ is, under general conditions, approximately normally distributed with mean Y and variance σ^2.

Thus, in particular, Proposition 2.2. states that $\hat{Y}_S$ is approximately unbiased and has the same approximate variance as $\hat{Y}_R$. This proposition is justified heuristically in the appendix, by an asymptotic result in Theorem 3.2 and by simulations in Section 4.

Remark 2.1. When all the p_i are equal, SPS is nothing but simple random sampling without replacement (srswor). In this case (2.7) reduces to the well-known formula for the variance of the srswor estimator of Y, except for a factor $(N-1)/N$. In order to ''calibrate'' σ^2 against this known ''standard'' one may multiply with a correction factor $N/(N-1)$,

$$\sigma^2 = \frac{1}{n}\frac{N}{(N-1)}\sum_{i=1}^{N}(1-np_i)\left(\frac{y_i}{p_i}-Y\right)^2 p_i \tag{2.10}$$

PS, on the other hand, is not equivalent to srswor in the equal probability case. However, its variance is approximately that of srswor, see Särndal et al. (1992, Equation 3.2.7). Hence, the $N/(N-1)$ correction of σ^2 may be proper in the PS case, too.

2.3. *Variation estimation*

For the sake of completeness we shall present variance estimators for PS and SPS, without going into theoretical details. Brewer and Hanif (1983, p. 83) suggest the following

''conventional estimator'' of the variance of the PS estimator.

$$\tilde{v}(\hat{Y}_R) = \frac{1}{n^2}\sum_{i\in s}(1-np_i)\left(\frac{y_i}{p_i} - \hat{Y}_R\right)^2 + \Pr(m=0)\hat{Y}_R^2 \tag{2.11}$$

Brewer and Hanif also state that ''a more stable estimator is obtained by multiplying the first expression on the right hand side by n/m.'' This is possible only if $m > 0$. We continue to assume that $\Pr(m = 0)$ is negligible and omit the right-most term in (2.11). By arguing as in Remark 2.1 we rather correct $v(\hat{Y}_R)$ by the quantity $n/(m-1)$), assuming $m > 1$ (which is necessary for variance estimation anyhow). This leads us to consider the following variance estimator for PS (which is left undefined for the case $m \leq 1$).

$$v(\hat{Y}_R) = \frac{1}{n(m-1)}\sum_{i\in s}(1-np_i)\left(\frac{y_i}{p_i} - \hat{Y}_R\right)^2 \tag{2.12}$$

The ''conventional estimator'' in case of SPS would be (2.11) without the right-most term. Again by a ''calibration to srswor'' argument the suggestion is to multiply this quantity by $n/(n-1)$. Thus, we arrive at the following variance estimator for SPS,

$$v(\hat{Y}_S) = \frac{1}{n(n-1)}\sum_{i\in s}(1-np_i)\left(\frac{y_i}{p_i} - \hat{Y}_s\right)^2 \tag{2.13}$$

which in the equal probability case reduces to the conventional, unbiased srswor variance estimator. The simulation studies give support for the use of the ''calibrated'' estimators of (2.12) and (2.13).

2.4. *Some notes on sampling in practice*

When performing SPS in practice we need not norm the auxiliary variable as in (1.1) since multiplication of p_i with a constant leaves the SPS sample unchanged. Hence, an SPS sample is drawn simply as follows. First pass through the file once to generate ξ_i. If the auxiliary variable is not normed as in (1.1), compute its sum S during this single pass of the file. Next sort the file in descending order of ξ_i. The first n units on the sorted list constitute the sample. By aid of S we check (2.1). The moving of a few units to a ''take-all'' stratum due to violation of (2.1) does not alter the sample, so we do not have to redraw the sample as with PS. The take-all units must of course be properly handled in the estimation process.

Because of the sorting, the procedure is neither *list-sequential* nor *draw-sequential* in the sense of Särndal et al. (1992, pp. 25–26). After sorting, SPS may be considered sequential in both respects, i.e., we draw a new unit by simply taking the next unit on the sorted list. With the list-sequential procedure PS we would have to pass through the whole list once more, which is inconvenient when sampling from a large register. This is the background to the name *sequential* PS.

Finally, some words on the application of SPS to the CPI. The retail trade section of Statistic Sweden's business register contains quite a lot of establishments that are out-of-scope for the CPI and cannot be detected from information in the register. With SPS it is possible to get a sample with a fixed number of in-scopes by simply excluding the out-of-scopes found in the sample and extending the sample with the next unit on the

list. By the independence of the involved random numbers, it is realized that the resulting ''net'' sample is just an SPS sample from an imaginary list of *in-scope* units. The traditional technique to achieve a fixed net sample, two-phase sampling, would be more expensive and is not suited for PRN coordination. Through PRN, the SPS samples for the Swedish CPI are not only coordinated over time but also with srswor samples for other business surveys. For more information about the CPI application of SPS we refer to Ohlsson (1990) and Dalén and Ohlsson (1995).

3. Asymptotic Results

Asymptotic results for finite populations require the introduction of a sequence of populations $\{U_k; k = 1, 2, 3, \ldots\}$, in which n and N tend to infinity. When referring to the kth member of this sequence we will add an index k to all quantities introduced in the previous sections. The first theorem is the asymptotic counterpart to Proposition 2.1. Let $\xrightarrow{d} N(0, 1)$ denote convergence in distribution to the standard normal distribution.

Theorem 3.1. Suppose Conditions (C1) and (C2) below are fulfilled. Then

$$\frac{\hat{Y}_{Rk} - Y_k}{\sigma_k} \xrightarrow{d} N(0, 1) \qquad \text{as } k \to \infty \tag{3.1}$$

The following is the asymptotic counterpart to Proposition 2.2.

Theorem 3.2. Suppose conditions (C1) and (C2) below are fulfilled. Then

$$\frac{\hat{Y}_{Sk} - Y_k}{\sigma_k} \xrightarrow{d} N(0, 1) \qquad \text{as } k \to \infty \tag{3.2}$$

The proofs of these theorems are given in Ohlsson (1995b). From Theorems 3.1 and 3.2 we conclude in particular that $\hat{Y}_S$ and $\hat{Y}_R$ are both asymptotically unbiased and that they are asymptotically equally efficient.

We now introduce the Conditions (C1) and (C2) previously referred to. Introduce the population mean $\bar{Y}_k = Y_k/N_k$ and define the following ''pps population variance''

$$\eta_k^2 = \sum_{i=1}^{N_k} \left(\frac{y_{ki}}{N_k p_{ki}} - \bar{Y}_k \right)^2 p_{ki} \tag{3.3}$$

The conditions are

$$\text{(C1)} \quad \frac{\max_i \left| \dfrac{y_{ki}}{N_k p_{ki}} - \bar{Y}_k \right|}{\sqrt{n_k}\, \eta_k} \to 0 \qquad \text{as } k \to \infty$$

(C2) There is a universal constant α, $0 < \alpha < 1$, such that for all i and k, $n_k p_{ki} \leq 1 - \alpha < 1$.

Remark 3.1. It is readily seen that (C1) implies $n_k \to \infty$, and hence $N_k \to \infty$. •

Remark 3.2. Since $1 = \sum_i p_{ki} \leq N_k \max_i \{p_{ki}\}$, (C2) implies $n_k/N_k \leq 1 - \alpha < 1$. •

3.1. Interpretation of the conditions

Despite the fact that Conditions (C1) and (C2) are *sufficient*, but not *necessary*, for (3.1) and (3.2) we believe that they give an idea of when the approximations of Propositions 2.1 and 2.2 are accurate.

We apply pps sampling when we believe in a strong relationship between the ys and ps. Condition (C1) indicates that we should avoid ''outliers'' (on a standardized scale) in this relationship in order for the approximations to be good. Since such outliers would increase the variance, the statement of Cochran (1977, p. 44) on srswor that ''good sampling practice tends to make the normal approximation more valid'' applies to PS and SPS, too.

Condition (C2) indicates that it might be a good idea to move units with inclusion probability very close to 1 to the take-all stratum. This recommendation is also supported by the simulation results in Section 4.

By Remark 3.1, we have an implicit condition that n should be large. In the PS case, n must in particular be large enough for $\Pr(m = 0)$ to be negligible.

4. Numerical Illustrations

Here we report results from two simulation (Monte Carlo) studies on PS and SPS. The object is to investigate the accuracy of the approximations in Propositions 2.1 and 2.2, to compare the precision of SPS to that of PS and to investigate the behaviour of the proposed variance estimators. The simulations were performed with the SAS system for PCs. The built-in random number generator RANUNI was used (with positive argument).

4.1. The CPI investigation of estimators and inclusion probabilities

Before the introduction of SPS in the Swedish CPI, its properties were investigated in a simulation study, parts of which were reported in Ohlsson (1990). Here we recapitulate these results.

Three populations were used in the study, represented by I–III below:

I. The total 1989 population of Swedish department stores, except for some out of scope units and the units which require inclusion probability 1. The measure of size, p, was the one actually used in the CPI, *viz.* number of employees plus one. The target value, Y, was a price index for a pair of men's socks from December 1987 to December 1988. This index is a weighted mean of price ratios for the stores, with weights p. Actual prices are of course only known for the CPI sample. For the other units, a value was imputed using the distribution of the known price changes. Since y is p times the price ratio, y and p are highly correlated here.

II. To get a very small population every tenth unit in *population I* was taken out to form *population II*.

III. A population with less correlation between y and p was found in the ''take-all'' part of the stratum ''manufacturing of machinery'' in Statistics Sweden's annual survey of financial accounts. The size p was again the number of employees plus one, while y was ''investments.''

Population I was investigated for two different sample sizes, $n = 41$, the actual size of

Table 1. Mean and standard deviation of the PS and SPS estimators

Set-up	1	2	3	4
Population	I	I	II	III
N	260	260	25	477
n	41	5	9	50
No. of iterations	6,000	1,500	7,020	6,000
$\Pr(m = 0)$	$3 \cdot 0^{-21}$	$6 \cdot 10^{-3}$	$8 \cdot 10^{-7}$	$6 \cdot 10^{-30}$
Y	106.93	106.93	106.46	18.822
Monte Carlo mean of $\hat{Y}_R$	106.93	106.00*	106.71	18.887
Monte Carlo mean of $\hat{Y}_S$	106.92	107.02	106.38	18.838
SD of srswr	12.00	34.36	25.96	6.10
SD of ppswr	3.34	9.55	9.22	3.85
SD of $\hat{Y}_{HT}$ from (2.5)	14.94	48.08	25.46	4.16
σ from (2.10)	2.99	9.45	7.01	3.52
Monte Carlo SD of $\hat{Y}_R$	3.03	14.26*	7.36	3.57
Monte Carlo SD of $\hat{Y}_S$	3.01	9.27	7.01	3.54

NOTE: *Estimator set to 0 for empty samples, in accordance with (2.6).

the 1989 CPI sample of department stores, and $n = 5$, about the smallest sample size in any CPI stratum. The other populations were examined with just one, somewhat arbitrary, choice of n each.

Table 1 contains the numerical results. For reference, several quantities that can be computed exactly are included, in particular the standard deviation (SD) of simple random sampling *with* replacement (*srswr*) and of pps sampling *with* replacement (*ppswr*). The relationship between these two figures may be used as an indicator of how well the population is suited for pps sampling.

In each step of the simulation, an independent set of random numbers was generated (the Xs in Section 2.1). This set was used to draw a PS sample and an SPS sample. The positive correlation due to the use of the same set of random numbers for both samples should, if anything, increase the precision in comparisons between the two. In Table 1, the mean and standard deviation over the simulations are reported as ''Monte Carlo mean'' and ''Monte Carlo SD.''

When comparing the output, it should be kept in mind that the number of iterations vary from set-up to set-up. The ''Monte Carlo'' standard error of the mean of $\hat{Y}_S$ is, e.g., only $3.01/\sqrt{6,000} = 0.039$ in set-up 1 but $9.27/\sqrt{1,500} = 0.239$ in set-up 2. In the latter case, 1,500 iterations were enough to illustrate the large difference between $\hat{Y}_S$ and $\hat{Y}_R$, while a larger number of iterations was considered necessary for the other set-ups.

By comparing the SD for ppswr (pps *with* replacement) to the value of σ, we find that in the realistic set-ups 1 and 4, we obtain around 10% variance reduction by using *without* replacement pps sampling. As expected, Table 1 shows that $\hat{Y}_{HT}$ is often of very poor precision.

Next we examine the simulation results for set-ups 1, 3, and 4. Here, both $\hat{Y}_R$ and $\hat{Y}_S$ are nearly unbiased and the bias is always a negligible part of their mean squared error. The approximation of the standard deviation by σ in (2.10) works very well for both $\hat{Y}_R$ and $\hat{Y}_S$. In all three cases, $\hat{Y}_S$ has the smallest standard deviation, but the difference to $\hat{Y}_R$ is small.

In set-up 2 we have a relatively high value of $\Pr(m = 0)$, and consequently some Poisson samples became empty here. The figures for $\hat{Y}_R$ are included for the sake of completeness, though one cannot recommend the use of Poisson sampling for such small n. For SPS, on the other hand, $\hat{Y}_S$ is almost unbiased and the Monte Carlo SD is quite close to σ, even with n as small as 5.

A trivial calculation shows that, if anything, the corrected version of σ, (2.10), performs better than the uncorrected (2.7).

The approximate unbiasedness of $\hat{Y}_S$ suggests that the SPS inclusion probabilities should be close to fulfilling (1.2), i.e., that SPS is approximately pps. This can be investigated directly by recording the relative frequency of inclusion in the simulated samples for every population unit. The entire results in all four set-ups is too extensive to report here. In Table 2 we give the desired and observed inclusion probabilities (inclusion frequencies) for a selection of the 260 units in set-up 1, ordered by size. 95% Monte Carlo confidence bounds were computed using the normal approximation to the binomial distribution.

We see that SPS is very close to strict pps with this set-up. The largest deviation (the only one being statistically significant at the 5% level) occurs for the largest unit, but is still very small. The other set-ups give the same picture, by and large. The largest deviation in any set-up was an observed probability of 96% where we wanted 91% (this difference

Table 2. Inclusion probabilities for SPS, set-up 1. Units sorted by size

ID	Desired probabilities	Observed probabilities		Confidence bounds
1	0.031	0.031	±	0.004
2	0.033	0.032	±	0.004
3	0.037	0.037	±	0.005
4	0.037	0.042	±	0.005
5	0.041	0.043	±	0.005
…				…
101	0.111	0.112	±	0.008
102	0.111	0.110	±	0.008
103	0.112	0.115	±	0.008
104	0.112	0.106	±	0.008
105	0.114	0.111	±	0.008
…				…
201	0.204	0.198	±	0.010
202	0.207	0.206	±	0.010
203	0.207	0.209	±	0.010
204	0.210	0.218	±	0.010
205	0.211	0.204	±	0.010
…				…
256	0.569	0.564	±	0.013
257	0.569	0.573	±	0.013
258	0.576	0.583	±	0.012
259	0.640	0.650	±	0.012
260	0.741	0.754	±	0.011

is highly significant with a Monte Carlo p-value of less than 10^{-10}). The fact that the largest deviations occur for the very largest units, together with Condition (C2), suggests that it may be a good idea to move a few units with desired inclusion probability close to 1 to the take-all stratum.

4.2. An investigation of estimators and variance estimators

Our next simulation study was performed in 1994, mainly to investigate the behaviour of the variance estimators $v(\hat{Y}_R)$ and $v(\hat{Y}_S)$, defined in (2.12) and (2.13), respectively. The study also gives further information on the mean and standard deviation of $\hat{Y}_R$ and $\hat{Y}_S$.

Here we use price changes from December 1991 to December 1992 for some fresh vegetables and pieces of furniture, as recorded by the Swedish CPI survey. The size measure p and target variable y are as in Section 4.1. (As seen in the tables, several prices actually decreased in 1992.) No imputations were made this time – our populations are the actual CPI samples.

The first four populations correspond to one CPI item each. The following, larger, populations were formed by putting together price quotations for different items two and two. For example, the population for set-up 5 below is the union of all price quotations for two vegetable items, treated here as prices of a single item.

In Table 3 we list the same quantities as in Table 1, plus the square root of the Monte Carlo mean of $v(\hat{Y}_R)$ and $v(\hat{Y}_S)$, and the Monte Carlo standard deviation of $v(\hat{Y}_R)$ and $v(\hat{Y}_S)$.

The conclusions from Table 1 remain valid. Furthermore, the square roots of the Monte Carlo means of both variance estimators are close to σ and to the Monte Carlo SDs of $\hat{Y}_R$ and $\hat{Y}_S$, respectively. It is not hard to see that the division by $(m-1)$ in (2.12) and $(n-1)$ in (2.13), rather than by m and n, reduces the bias. Finally, $v(\hat{Y}_R)$ and $v(\hat{Y}_S)$ are approximately equally stable, as measured by their Monte Carlo SDs. If anything, the latter is more stable.

5. Conclusions

In our opinion, simplicity is a virtue of a sampling procedure. Both PS and SPS are very simple to use in practice; with computer packages, they typically require even less programming than pps with replacement. With PRN, either procedure gives a simple solution to the problem of updating pps samples while retaining a large number of units, and to other sample coordination problems. When using pps with replacement, one could repeatedly use the Kish and Scott (1971) method for updating $n = 1$ samples. This method is much more complicated than the PRN approach, though, and cannot be used for controlling overlap with other surveys. Hence, even in cases where the variance reduction of using without replacement sampling is small, PS and SPS are preferable to with replacement sampling when we are conducting a repeated survey.

PS has the merit over SPS of being *strictly* pps. One may doubt the value of this merit, though, since the estimator based on this fact, the Horvitz-Thompson estimator, is very inefficient for PS. Conditionally on the sample size, Poisson is no longer strictly pps. For the ratio-type estimator that is recommended for use with PS we have to rely on approximate results for the mean and variance, just as is the case with SPS. We have shown that the SPS estimator and the PS ratio estimator are asymptotically equally

Table 3. Mean and standard deviation of PS and SPS estimators and variance estimators. SD = standard deviation

Set-up	1	2	3	4	5	6	7	8
Item type	Vegetab.	Vegetab.	Furniture	Furniture	Vegetab.	Vegetab.	Furniture	Furniture
N	63	55	48	48	124	116	96	94
n	10	10	10	10	20	20	22	22
No. of iterations	1,500	1,500	2,218	1,500	1,500	1,500	1,500	1,500
$\Pr(m = 0)$	$3 \cdot 10^{-6}$	$1 \cdot 10^{-6}$	$7 \cdot 10^{-7}$	$7 \cdot 10^{-7}$	$3 \cdot 10^{-6}$	$1 \cdot 10^{-6}$	$7 \cdot 10^{-7}$	$7 \cdot 10^{-7}$
Y	81.02	97.25	104.78	105.37	89.45	95.79	105.08	99.04
Monte Carlo mean of $\hat{Y}_R$	80.88	97.14	104.86	105.36	89.39	95.86	105.10	99.07
Monte Carlo mean of $\hat{Y}_S$	80.93	97.12	104.86	105.42	89.27	95.86	105.10	99.03
SD of srswr	30.22	33.05	37.39	36.19	22.96	23.71	24.55	22.38
SD of ppswr	12.38	7.03	3.90	3.93	7.23	5.48	2.64	3.00
SD of $\hat{Y}_{HT}$ from (2.5)	23.72	25.06	24.48	25.04	17.40	17.69	15.99	15.27
σ from (2.10)	10.74	6.09	3.29	3.25	6.20	4.50	2.12	2.20
Monte Carlo SD of $\hat{Y}_R$	11.20	6.05	3.28	3.28	6.26	4.68	2.12	2.22
Monte Carlo SD of $\hat{Y}_S$	10.91	5.93	3.19	3.18	6.09	4.64	2.11	2.20
$\sqrt{}$ Monte Carlo mean of $v(\hat{Y}_R)$	10.54	5.99	3.19	3.17	6.17	4.48	2.10	2.19
$\sqrt{}$ Monte Carlo mean of $v(\hat{Y}_S)$	10.54	6.02	3.21	3.19	6.17	4.49	2.11	2.20
Monte Carlo SD of $v(\hat{Y}_R)$	113.42	19.02	9.98	10.11	28.50	6.37	2.90	1.47
Monte Carlo SD of $v(\hat{Y}_S)$	107.99	18.44	9.80	9.98	27.96	6.30	2.89	1.48

efficient. The simulations also show that there is little to choose between these estimators on efficiency grounds.

Being almost equal in the above respects, the choice between PS and SPS must fall on SPS because of its fixed sample size, and (for some applications) because of the possibility of sampling a fixed number of in-scope units. The fixed size allows us to use SPS in cases where PS is at risk of giving an empty sample, e.g., when $n = 5$. The properties of SPS for small samples ($n < 5$, say) remain to be investigated, though.

Appendix: Heuristics Proofs

Here we give heuristic arguments for the approximation results Propositions 2.1 and 2.2 in a ''finite'' (single population) situation. The formal proofs of the corresponding asymptotic results, Theorems 3.1 and 3.2, are given in Ohlsson (1995b).

We first define an auxiliary stochastic process $\{Z(t), t \geq 0\}$.

$$Z(t) = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{p_i} - Y\right)1\{\xi_i \leq t\} \tag{A.1}$$

where $1\{\cdot\}$ denotes the indicator function. Again, we neglect the possibility $m = 0$ in the Poisson case, assuming n to be large. An ordinary Taylor linearization of the ratio $\hat{Y}_R$ now yields

$$\hat{Y}_R - Y = \frac{n\hat{Y}_{HT} - mY}{m} \approx \hat{Y}_{HT} - \frac{m}{n}Y = Z(n) \tag{A.2}$$

The summation in $Z(t)$ will get non-zero contributions from objects with $\xi_i \leq t$, i.e., for units in a Poisson sample of expected size t. Let T be the lowest value of t for which we get exactly n units in this sample. Then $T = \xi(n)$, the nth smallest of the ξ. From (A.1) and (2.9) we see that $\hat{Y}_S - Y = Z(T)$. Viewed this way, sequential Poisson sampling is a Poisson sample for which we have adjusted the expected size to get a sample of exactly size n. The probability distribution of T will intuitively be centered around n. This motivates the approximation

$$\hat{Y}_S - Y = Z(T) \approx Z(n) \tag{A.3}$$

Giving an asymptotic result corresponding to the approximation in (A.3) is the core of the stringent proof in Ohlsson (1995b). The proof relies on results in Rosén (1997).

Having established the approximations (A.2) and (A.3), the rest is an easy task. It is readily seen that $Z(n)$ has zero mean and variance σ^2 given by (2.7). Furthermore, $Z(n)$ is a sum of independent random variables and is thus approximately normally distributed, by the Lindeberg-Liapunov central limit theorem. In Ohlsson (1995b), the Conditions (C1) and (C2) are shown to be sufficient for this theorem to be valid.

6. References

Atmer, J., Thulin, G., and Bäcklund, S. (1975). Samordning av urval med JALES-metoden. Statistisk tidskrift, 13, 443–450. (In Swedish with English summary.)

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972). Selecting Several Samples From a Single Population. Australian Journal of Statistics, 14, 231–239.

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

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. New York: Wiley.

Dalén, J. and Ohlsson, E. (1995). Variance Estimation in the Swedish Consumer Price Index. Journal of Business and Economic Statistics, 13, 347–356.

Feller, W. (1950). An Introduction to Probability Theory and Its Applications. New York: Wiley.

Hájek, J. (1960). Limiting Distributions in Simple Random Sampling from a Finite Population. Publications of the Mathematical Institute of the Hungarian Academy of Science, 5, 361–371.

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

Keyfitz, N. (1951). Sampling with Probabilities Proportional to Size. Journal of the American Statistical Association, 46, 105–109.

Kish, L. and Scott, A. (1971). Retaining Units After Changing Strata and Probabilities. Journal of the American Statistical Association, 66, 461–470.

Ogus, J.L. and Clark, D.F. (1971). The Annual Survey of Manufacturers: A Report on Methodology. Technical Paper No. 24. U.S. Bureau of the Census, Washington, DC.

Ohlsson, E. (1990). Sequential Poisson Sampling from a Business Register and Its Applications to the Swedish Consumer Price Index. R&D Report 1990:6, Statistics Sweden, Stockholm.

Ohlsson, E. (1995a). Coordination of Samples Using Permanent Random Numbers. In Cox, B. et al. (eds.): Business Survey Methods. New York: Wiley, 153–169.

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

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

Särndal, C.E. (1996). Efficient Estimators with Simple Variance in Unequal Probability Sampling. Journal of the American Statistical Association, 91, 1289–1300.

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

Sigman, R.S. and Monsour, N.J. (1995). Selecting Samples from List Frames of Businesses. In Cox, B. et al. (eds.): Business Survey Methods. New York: Wiley, 133–152.

Sunter, A.B. (1977). Response Burden, Sample Rotation, and Classification Renewal in Economic Surveys. International Statistical Review, 45, 209–222.

Sunter, A.B. (1989). Updating Size Measures in a PPSWOR Design. Survey Methodology, 15, 253–260.

Templeton, R. (1990). Poisson Meets the New Zealand Business Directory. The New Zealand Statistician, 25, 2–9.

Received July 1996
Revised April 1997