Survey Methodology, June 1996
Vol. 22, No. 1, pp. 85–94
Statistics Canada

A Moving Stratification Algorithm

YVES TILLÉ[1]

ABSTRACT

A general algorithm with equal probabilities is presented. The author provides the second order inclusion probabilities that correspond to the algorithm, which generalizes the selection-rejection method, so that a sample may be drawn using simple random sampling without replacement. Another particular case of the algorithm, called moving stratification algorithm, is discussed. A smooth stratification effect can be obtained by using, as a stratification variable, the serial number of the observation units. The author provides approximations of first and second order inclusion probabilities. These approximations lead to a population mean estimator and to an estimator of the variance of this mean estimator. The algorithm is then compared to a classical stratified plan with proportional allocation.

KEY WORDS: Selection algorithm; Equal probability sampling; Strata.

1. INTRODUCTION

When a file is ordered according to an auxiliary variable that is close to the variable of interest, how can a sample be selected using such information? One solution to the problem consists of making a stratified selection. However, making such a selection requires that a delicate problem be resolved, namely subdividing the population into strata. Another simple solution that is both quick and efficient consists of making a systematic selection. The algorithm can be written in a few lines. Moreover, the way in which the file is ordered can be put to good use. However, a systematic selection has one major flaw, namely that estimating the variance of total or mean estimators requires one or several hypotheses concerning the population. It will be shown that there is another simple selection algorithm with which a sample can be drawn in one pass using the file ordering system. For this algorithm, an estimator of the variance of a total or mean estimator is provided, requiring no modelling of the population.

A general selection algorithm providing equal first order inclusion probabilities is presented in section 2. First and second order inclusion probabilities are provided. In section 3, the proposed algorithm is shown to generalize the selection-rejection method so that a simple random sample can be drawn without replacement along with the stratified plan with proportional allocation. Finally, in section 4, the moving stratum method is defined and, in section 5, conclusions are drawn.

2. PRESENTATION OF THE GENERAL ALGORITHM

2.1 The Algorithm

Let us consider a finite population $U = \{1, \ldots, i, \ldots, N\}$; we write $y_1, \ldots, y_i, \ldots, y_N$, the N values assumed by variable y for N observation units of U. The mean of the values assumed by variable y for the population is written as

$$\bar{y} = \frac{1}{N} \sum_{i \in U} y_i.$$

A random sample s of fixed size n is drawn from this population. The random variables indicating the presence of observation units in s are written as I_i, $i \in U$. The first order inclusion probability is written as $\pi_i = \Pr(i \in s) = E(I_i)$, $i \in U$ and the second order inclusion probability as $\pi_{ik} = E(I_i I_k)$, $i \neq k \in U$. The algorithm is very short. It resembles the algorithms of Fan, Fuller and Rezucha (1962), Bebbington (1975), McLeod and Bellhouse (1983) and Sunter (1977, 1986). Only N, n and the b_i, $i = 0, \ldots, N - 1$ need to be known. The other variables are working variables.

General Algorithm

```
$j <= 0$;
$i <= 0$;
Repeat for $i = 0, \ldots, N - 1$
    $u <=$ a random number with a uniform distribution [0,1];
        if $\frac{(b_i + i)n/N - j}{b_i} > u$ then
            select record $i + 1$;
            $j <= j + 1$;
        otherwise, pass the record $i + 1$;
    $i <= i + 1$.
```

At each step, j represents the number of records already selected and i the number of records passed (selected or not). For each iteration, a decision is made about selecting the record $i + 1$. If the record is selected, it becomes the $(j + 1)$-th in the sample. The coefficients b_i, $i = 0, \ldots, N - 1$, are strictly positive real numbers. These

[1] Yves Tillé, Laboratoire de Méthodologie du Traitement des Données, C.P. 124, Université Libre de Bruxelles, avenue Jeanne, 44, 1050 Bruxelles, Belgique, E-mail ytilleb@ulb.ac.be

quantities must meet certain conditions discussed below if the plan is to be of fixed size or if the units are to be selected with equal probability. The choice of different values for b_i, $i = 0, \ldots, N - 1$, will make it possible to generate several special cases of the general algorithm.

If b_i are strictly positive reals such that $b_i \leq N - i$, then the sample size is equal to or smaller than n. In fact, assuming we have already drawn n units from the population at step i and that $b_i \leq N - i$, then

$$\frac{(b_i + i)n/N - n}{b_i} = \frac{n}{N} - \frac{n}{b_i}\frac{N-i}{N} \leq \frac{n}{N} - \frac{n}{N-i}\frac{N-i}{N} = 0.$$

It becomes impossible to draw a further unit. It will be assumed in everything that follows that $b_i \leq N - i$. Moreover, if $b_i \leq N - i$, $i = 1, \ldots, N - n - 1$ and if $b_i = N - i$, $i = N - n, \ldots, N - 1$, the sample is of fixed size n. Note that these conditions for obtaining a sample of fixed size are sufficient but not necessary.

Three particular cases of the algorithm are examined below. These three cases are defined by three choices of coefficient b_i, $i = 0, \ldots, N - 1$. Before examining these particular choices, we will determine the first and second order inclusion probabilities without loss of generality.

2.2 First Order Inclusion Probabilities

We write n_i, the number of units selected after passing i records. We see immediately that $n_1, \ldots, n_i, \ldots, n_N$ is a Markov chain. In fact, we directly derive from the algorithm that

$$\Pr[n_i = j \mid n_1, \ldots, n_{i-1}] = \Pr[n_i = j \mid n_{i-1}].$$

The random variables

$$c_i = \frac{(b_i + i)n/N - n_i}{b_i}, \; i = 0, \ldots, N - 1,$$

can sometimes assume values greater than 1 or less than 0. Since $\max(0, n - N + i) \leq n_i \leq \min(i,n)$, then $\Pr[0 \leq c_i \leq 1] = 1$ if

$$b_i \geq \begin{cases} \min\left(i\dfrac{N-n}{n}, N - i\right) & \text{if } n \leq N/2 \\ \min\left(i\dfrac{n}{N-n}, N - i\right) & \text{if } n > N/2 \end{cases},$$

$$i = 0, \ldots, N - 1. \quad (1)$$

Again conditions (1) are sufficient but not necessary. We can therefore construct b_i which do not meet these conditions but which provide c_i in $[0,1]$. The case dealt with in section 3.2 (stratification) represents one example.

The following example also provides c_i in $[0,1]$ without meeting condition (1): let us consider $N = 12$, $n = 4$ and $b_0 = b_1 = b_3 = b_4 = b_6 = 6$, $b_2 = b_5 = 7$, $b_i = N - i, i = 12 - i, i = 7, \ldots, 11$. We have $c_0 = 1/3$, $c_1 = (7 - 3n_1)/18$, $c_2 = (3 - n_2)/7$, $c_3 = (3 - n_3)/6$, $c_4 = (10 - 3n_4)/18$, $c_5 = (4 - n_5)/7$, $c_6 = (4 - n_6)/6$, $c_7 = (4 - n_7)/5$, $c_8 = (4 - n_8)/4$, $c_9 = (4 - n_9)/3$, $c_{10} = (4 - n_{10})/2$, $c_{11} = (4 - n_{11})$. We note that $n_1 \leq 1$, $n_2 \leq 2$, $n_3 \leq 3$. If $n_3 = 3$ then $c_3 = 0$ and therefore $n_4 \leq 3$. We then have $n_5 \leq 4$ and if $n_5 = 4$ then $c_5 = 0$ and therefore $n_6 \leq 4$. This last comment is true for all c_i that follow. We therefore note that all c_i are in $[0,1]$ whereas $b_4 = 6$ does not meet condition (1).

In order to simplify the demonstrations which follow, it will be assumed that

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N - 1.$$

We will return to the problem of c_i values greater than 1 or smaller than 0 later on. If

$$\Pr[0 \leq c_i \leq 1] = 1, \; i = 0, \ldots, N - 1,$$

we have

$$E[I_{i+1} \mid n_1, \ldots, n_i] = E[I_{i+1} \mid n_i] = \frac{(b_i + i)n/N - n_i}{b_i}.$$

It can be shown easily by recursion that if $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N - 1$, $E[n_i] = i\,n/N$, $i = 0, \ldots, N$. Therefore,

$$\pi_i = E[I_i] = E[n_i] - E[n_{i-1}] = \frac{n}{N}. \quad (2)$$

2.3 Second Order Inclusion Probabilities

Four results provided by lemmas 1, 2 and 3 are needed in order to determine second order inclusion probabilities.

Lemma 1 If $\Pr[0 \leq c_i \leq 1] = 1$, $i = 0, \ldots, N - 1$, then

$$E[n_{i+k} \mid n_i]$$

$$= (i + k)\frac{n}{N} + \left(n_i - i\frac{n}{N}\right)\prod_{\ell=i}^{i+k-1}\frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \ldots, N - 1, k = 1, \ldots, N - i.$$

This lemma can be demonstrated by recursion if it is assumed to be true for $k - 1$. Using lemma 1, the following lemma is readily obtained by subtraction:

Lemma 2 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$E[I_{i+k} \mid n_i] = \frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_{i+k-1}} \prod_{\ell=i}^{i+k-2} \frac{b_\ell - 1}{b_\ell},$$

$$i = 1, \ldots, N-1, k = 1, \ldots, N-i.$$

It is assumed by convention that an empty product has a value of 1.

Lemma 3 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$\operatorname{Var}[n_i] = \frac{n}{N}\frac{N-n}{N} \sum_{j=1}^{i} \prod_{\ell=j}^{i-1} \frac{b_\ell - 2}{b_\ell}, i = 1, \ldots, N. \quad (3)$$

The demonstration is provided in the appendix.

Finally, the second order inclusion probability is provided by the following proposition:

Proposition 1 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$E[I_{i+k}I_{i+1}] = \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\frac{1}{b_{i+k-1}}$$

$$\times \left(1 - \frac{1}{b_i}\sum_{j=1}^{i}\prod_{\ell=j}^{i-1}\frac{b_\ell - 2}{b_\ell}\right)\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell},$$

$$i = 0, \ldots, N-2, k = 2, \ldots, N-i. \quad (4)$$

The demonstration is provided in the appendix.

Corollary 1 If $\Pr[0 \le c_i \le 1] = 1,\ i = 0, \ldots, N-1$, then

$$\pi_{ik} = \frac{n^2}{N^2} - \frac{n}{N}\frac{N-n}{N}\left(1 - \frac{1}{b_{i-1}}\sum_{j=1}^{i-1}\prod_{\ell=j}^{i-2}\frac{b_\ell - 2}{b_\ell}\right)$$

$$\times \frac{1}{b_{k-1}}\prod_{\ell=i}^{k-2}\frac{b_\ell - 1}{b_\ell}, \ i = 1, \ldots, N-1, k > i.$$

2.4 The Horvitz-Thompson Estimator and its Variance

The Horvitz-Thompson estimator is the simple sample mean since the first order inclusion probabilities are all equal

$$\hat{\bar{y}}_\pi = \frac{1}{n}\sum_{i \in s} y_i.$$

If the design is of fixed size, we can use the Yates and Grundy variance formula (1953)

$$\operatorname{Var}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i \in U}\sum_{\substack{k \in U \\ k \ne i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2(\pi_i\pi_k - \pi_{ik}). \quad (5)$$

Since $\pi_i = n/N, i = 1, \ldots, N$ and assuming that

$$\gamma_{ik} = 1 - \pi_{ik}\frac{N^2}{n^2},$$

we can write

$$\operatorname{Var}[\hat{\bar{y}}_\pi] = \frac{1}{N^2}\sum_{i \in U}\sum_{\substack{k \in U \\ k \ne i}}(y_i - y_k)^2\gamma_{ik}. \quad (6)$$

The variance estimator is provided by

$$\widehat{\operatorname{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2N^2}\sum_{i \in s}\sum_{\substack{k \in s \\ k \ne i}}\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2\frac{\pi_i\pi_k - \pi_{ik}}{\pi_{ik}}. \quad (7)$$

This can be written here as

$$\widehat{\operatorname{Var}}[\hat{\bar{y}}_\pi] = \frac{1}{2n^2}\sum_{i \in s}\sum_{\substack{k \in s \\ k \ne i}}(y_i - y_k)^2\frac{\gamma_{ik}}{1 - \gamma_{ik}}.$$

3. APPLICATION 1: SIMPLE AND STRATIFIED RANDOM SELECTIONS

3.1 Simple Design

The simplest selection algorithm, the selection-rejection method described in Fan, Fuller and Rezucha (1962, method 1), Beddington (1975) and Deville and Grosbras (1987, p. 210), is of course a particular case of the general algorithm. We need only take

$$b_i = N - i, \quad i = 0, \ldots, N-1.$$

We always have $0 \le c_i \le 1$. The first order inclusion probabilities always have a value of n/N. Calculations for second order inclusion probabilities follow from proposition 1. Assuming $k > i$, on the basis of corollary 1, we can find the second order inclusion probabilities of the simple design:

$$\pi_{ik} = \frac{n(n-1)}{N(N-1)}.$$

We also recall some classical results concerning the simple design that we will be using later on. The estimator for $\bar{y}$ is therefore the mean of the sample

$$\hat{\bar{y}}_{srs} = \frac{1}{n} \sum_{i \in s} y_i. \tag{8}$$

The variance of this estimator is provided by

$$\mathrm{Var}\left[\hat{\bar{y}}_{srs}\right] = \frac{\sigma_y^2}{n} \frac{N - n}{N - 1} \tag{9}$$

where

$$\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{y})^2. \tag{10}$$

An unbiased estimate of this variance is

$$\widehat{\mathrm{Var}}\left[\hat{\bar{y}}_{srs}\right] = \frac{s_y^2}{n} \frac{N - n}{N} \tag{11}$$

where

$$s_y^2 = \frac{1}{n - 1} \sum_{i \in s} \left(y_i - \hat{\bar{y}}_{srs}\right)^2. \tag{12}$$

3.2 Stratified design

The stratified design can also be defined using the general algorithm. The stratification variable in this case is the serial number of the individual. Let us consider the particular case of a stratified design of H strata with proportional allocation where all the strata are of the same size. The strata are such that the individuals of a given stratum are adjacent in the data file. It is also assumed that N/H is an integer. This stratified design is obtained by simply taking

$$b_i = \left\{(N - i - 1) \bmod \frac{N}{H}\right\} + 1, \; i = 0, \ldots, N - 1.$$

4. APPLICATION 2: MOVING STRATIFICATION

4.1 The Problem

The file is assumed to be ordered according to an auxiliary variable that is close to the variable of interest. The problem is as follows: how can we draw a random selection that yields a small variance for the Horvitz-Thompson estimator of a mean? Looking at the formulation of the Yates-Grundy variance (5), we see that there are two distinct answers to this question.

The first solution consists of selecting with unequal probabilities using first order inclusion probabilities that are proportional to the variable of interest. If such a selection could be made, all quantities

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2$$

would be zero and therefore the variance would be zero.

The second solution consists of using second order inclusion probabilities. A good selection could be one where π_{ik} are close to $\pi_i \pi_k$ if y_i is very different from y_k. On the other hand, if y_i is very close to y_k, we can select second order inclusion probabilities π_{ik} that are clearly smaller than $\pi_i \pi_k$. Thus, where quantities

$$\left(\frac{y_i}{\pi_i} - \frac{y_k}{\pi_k}\right)^2$$

would be large (respectively small), quantities $\pi_i \pi_k - \pi_{ik}$ would be small (respectively large). We would thus have a small variance.

The second solution we have just described is in fact often used. It is the basic idea for stratification. Our objective is to apply this idea to the construction of a sequential selection algorithm that is easy to implement. Such an algorithm could be applied to any file without the need to know anything save the size of the population. It would therefore apply to very large files. We could thus benefit from the information provided by this auxiliary variable like for stratification, without the need to actually subdivide into strata.

4.2 The Method

We first define M the length of the moving stratum within the population. M represents, in a way, the size of the stratum within the population and is such that $N/n \leq M \leq N$. The algorithm of the moving stratum is defined by

$$b_i = \min(M, N - i), \; i = 0, \ldots, N - 1.$$

There is, however, one problem. Quantities c_i defined by

$$c_i = \begin{cases} \dfrac{(M + i)n/N - n_i}{M} & \text{if } i \leq N - M \\ \dfrac{n - n_i}{N - i} & \text{otherwise,} \end{cases}$$

are not always in $[0,1]$.

In fact, let us assume that, before the $(N - M)$-th step of the algorithm, c_i is positive and very close to zero and that through some bad luck the unit i is nevertheless chosen. In such a case, c_{i+1} would have a value of $c_i - (N-n)/(NM)$. c_{i+1} can therefore have a negative value but this negative value is always greater than $-(N-n)/(NM)$. In fact, if one of the c_i is already negative, the unit i is not selected and therefore c_{i+1} has a value greater than c_i.

Let us now assume that before the $(N - M)$-th step of the algorithm, one c_i is very slightly smaller than 1 and that nevertheless unit i is not selected. In such a case, c_{i+1} would have a value of $c_i + n/(NM)$. c_{i+1} can therefore take on a value greater than 1 but this value greater than 1 is nevertheless always smaller than $1 + n/(NM)$. In fact, if one of the c_i is already greater than 1, the unit i is always selected and therefore c_{i+1} has a value smaller than c_i.

We obtain

$$\Pr\left[-\frac{N-n}{NM} < c_i < 1 + \frac{n}{NM}\right] = 1, i = 0, \ldots, N - M. \tag{13}$$

The design is however of fixed size, a result that follows the following proposition:

Proposition 2 If $b_i = \min(M, N - i)$, $(N/n < M < N)$, $0 = 1, \ldots, N - 1$, then the design is of fixed size.

The demonstration is provided in the appendix.

Since the c_i are not always within the interval $[0,1]$, we carried out 50 simulations of the moving stratum algorithm for various sample and population sizes. The selected N population sizes were 100, 500, 2500, 12500, 62500, 312500. The reciprocals of sampling rates (N/n) were 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. We carried out several simulations by varying the size of the moving stratum as follows: $M = N/n, 2N/n, 3N/n, \ldots$. The simulations seem to indicate that the greater the value for M, the smaller the probability that a c_i will fall outside of $[0,1]$. As soon as $M \geq 10N/n$, for all the simulations that we carried out, the problem was no longer raised. This first result does not imply that the probability that at least one of the c_i will fall outside of $[0,1]$ is zero when $M \geq 10N/n$. However, it may be said that such a probability would then be very small.

4.3 Estimating the Mean and Bias

In examining the results yielded by expression (2) and proposition 1, we get, as a first approximation, a value of about $\pi_i \approx n/N$ for first order inclusion probabilities. This approximation of inclusion probabilities makes it possible to construct an estimator.

$$\hat{\bar{y}}_{sm} = \frac{1}{n} \sum_{i \in s} y_i.$$

This estimator is slightly biased since the c_i are not all exactly within the interval $[0,1]$. This bias is

$$B\left[\hat{\bar{y}}_{sm}\right] = \frac{1}{N} \sum_{i \in U} \alpha_i y_i$$

where $\alpha_i = \pi_i N/n - 1$. Since the design is of fixed size, $\sum_{i \in U} \alpha_i = 0$. We can therefore write the bias in the form of a covariance: $B\left[\hat{\bar{y}}_{sm}\right] = \sigma_{y\alpha}$ where

$$\sigma_{y\alpha} = \frac{1}{N} \sum_{i \in U} \alpha_i (y_i - \bar{y}). \tag{14}$$

Since the absolute value of a covariance is always equal to or smaller than the product of the two standard deviations, we obtain an upper bound for the absolute value of the bias

$$\left| B\left[\hat{\bar{y}}_{sm}\right] \right| \leq \sigma_y \sigma_\alpha$$

where σ_y is defined by (10) and

$$\sigma_\alpha^2 = \frac{1}{N} \sum_{i \in U} \alpha_i^2.$$

The variance of the estimator is of a magnitude that is comparable (for N and fixed n) to the variance of the estimator of the mean in the simple design without replacement. We can therefore write

$$\left| B\left[\hat{\bar{y}}_{sm}\right] \right| \leq C_\alpha \sqrt{\operatorname{Var}\left[\hat{\bar{y}}_{srs}\right]}$$

where Var $\left[\hat{\bar{y}}_{srs}\right]$ is defined by (9) and

$$C_\alpha = \sigma_\alpha \sqrt{\frac{n(N-1)}{(N-n)}}.$$

We will assume that the bias is negligible when the upper bound of the bias of the estimator $\hat{\bar{y}}_{sm}$ is negligible with respect to $\operatorname{Var}\left[\hat{\bar{y}}_{srs}\right]^{1/2}$, *i.e.*, when C_α is small.

Recursively we can calculate the exact value of the $\Pr[n_i = j]$ since we have

$$\Pr[I_i = 1 \mid n_i] = \tilde{c}_i, i = 1, \ldots, N - M$$

where $\tilde{c}_i$ has a value of 0 if $c_i < 0$, c_i if $0 \leq c_i \leq 1$ and 1 if $c_i > 1$. From this result we can derive the exact value of first order inclusion probabilities.

We have calculated (Appendix, Table 1) the values of C_α for various sample and population (100 - 312500) sizes. The values of C_α are provided for sizes of moving strata M equal to N/n, $2N/n$, $3N/n$, $4N/n$ and $5N/n$. It can be seen that as soon as the value of the moving stratum is $2N/n$, C_α never exceeds 0.07. When $M = 3N/n$, the coefficient C_α is expressed in thousandths. According to Cochran (1977, pp. 13-14), the bias is then negligible. The table therefore shows that if $M \geq 3N/n$, the bias of the estimator will be negligible at least for the specified sample and population sizes.

However, these results do not imply that the bias of the estimator is large when M is very small (for example $M = N/n$). The C_α are bias upper bounds. From expression (14), we see that the bias will be all the greater as the variable of interest correlates with the exact inclusion probabilities. We have shown (Figure 1) the exact inclusion probabilities (y axis) for N individuals (x axis) obtained by using the moving stratification algorithm with the

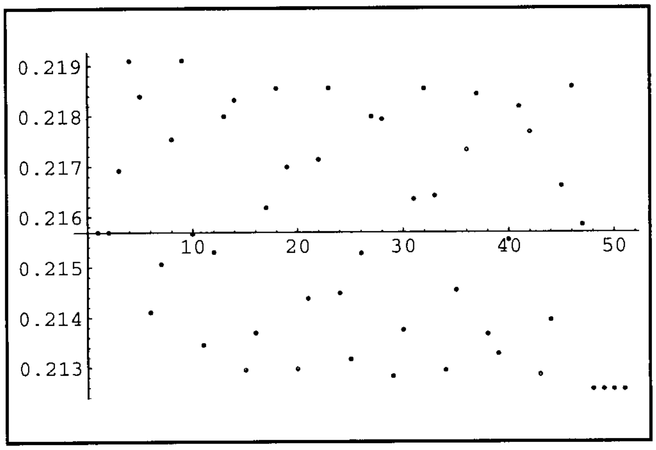

Figure 1. Inclusion probabilities.

parameters $N = 51$, $n = 11$, $M = N/n$. This case is obviously very unfavourable. The result is interesting. In this case, $n/N = 0.215686$. The inclusion probabilities are distributed on both sides of n/N with no marked tendency associated with the ordering of the file. In practical terms, the probability can be considered very small that there will be a variable of interest that strongly correlates with the exact inclusion probabilities; as a result, the bias will most often be clearly smaller than the given upper bound.

We could, of course, use the exact inclusion probabilities to establish an estimate. We feel that this is not worthwhile, for two reasons:

- first, because calculating the exact inclusion probabilities requires a significant amount of time,
- second, because the exact first order inclusion probabilities are such that

$$\mathrm{Var}\left[\sum_{i\in s}\frac{1}{\pi_i}\right] \neq 0.$$

In this case, we have a random Horvitz-Thompson estimator of a constant variable ($y_k = C$). To overcome this problem, an estimate of the mean is usually carried out using Hájek's (1971) ratio. This estimator is also biased.

4.4 Estimating the Variance of the Estimator

Assuming that $\Pr(0 \leq c_i \leq 1) \approx 1$, we can also build an approximation of second order inclusion probabilities using corollary 1. Given that b_i has a value of M if $i \leq N - M$ and $N - i$ otherwise, we obtain the following approximation:

$$\pi_{ik} \approx \frac{n^2}{N^2}(1 - \theta_{ik})$$

where

$$\theta_{ik} = \frac{N-n}{2n}\frac{1}{M-1}\left\{1 + \left(\frac{M-2}{M}\right)^{\min(i-1,N-M)}\right\} \times \left(\frac{M-1}{M}\right)^{\max(0,\min(N-M-i+1,k-i))} \quad k > i.$$

Assuming that the first order inclusion probabilities have a value of n/N, an approximation of the variance of $\hat{\bar{y}}_{sm}$ can be obtained:

$$\mathrm{Var}_{app}\left[\hat{\bar{y}}_{sm}\right] = \frac{1}{2N^2}\sum_{i\in U}\sum_{\substack{k\in U\\k\neq i}}(y_i - y_k)^2\theta_{ik}. \qquad (15)$$

From (15), an estimator of the variance of the estimator of the mean can be obtained:

$$\widehat{\mathrm{Var}}_{app}\left[\hat{\bar{y}}_{sm}\right] = \frac{1}{2N^2}\sum_{i\in s}\sum_{\substack{k\in s\\k\neq i}}(y_i - y_k)^2\frac{\theta_{ik}}{1-\theta_{ik}}. \qquad (16)$$

Again, this estimator is biased. In order to assess the magnitude of the bias, we carried out a series of simulations. The results are given in Table 2 in the appendix. We generated populations of size $N = 400$. The values assumed by the two variables x and y were generated by means of pseudo-random numbers having a bivariate normal distribution with a fixed coefficient of correlation ρ. The populations were then sorted in terms of the variable x. The objective was to estimate $\bar{y}$.

In these populations, samples of size 64 were selected using the moving stratum method (*sm*), a stratified design with proportional allocation in which the sizes of the strata were all equal (*strat*), as well as a simple design without replacement (*srs*). These three methods are particular cases of the general algorithm and they were implemented using the same random numbers. Simulations were carried out for different values of the moving stratum M (case: *sm*) and for different numbers of strata H (case: *strat*). An explanation is provided below for the choices of M and H. For each simulation, 200,000 samples were selected.

For each of the simulations, three results are given:

- The means for the simulations of the estimators of the variance of the estimator of the mean, which are expressed as $E_{sim}\widehat{\mathrm{Var}}(\hat{\bar{y}})$. These variance estimators are given by expressions (11) (*srs*) and (16) (*sm*).
- The mean-square errors for the simulations of the estimators of the mean. These quantities are expressed as $EQM_{sim}(\hat{\bar{y}}) = E_{sim}(\hat{\bar{y}} - \bar{y})^2$.
- The variances of the estimators of the mean. These variances are given by expressions (9) (*srs*) and (15) (*sm*). In the case of the moving stratification, this is of course the proposed approximation.

A careful reading of the results seems to indicate that the variance estimator proposed for the moving stratum algorithm is not affected by a systematic bias no matter what the value for the coefficient of correlation between x and y. The results also seem to indicate that the approximate expression given for the variance of the estimator of the mean for the moving stratification is a valid approximation.

4.5 Interest of the Algorithm

Within the class of algorithms defined by the general algorithm, we call the mean horizon of an algorithm the quantity

$$\bar{b} = \frac{1}{N} \sum_{i=0}^{N-1} b_i.$$

For the simple design, we get $\bar{b}_{srs} = (N + 1)/2$. For the algorithm of the moving stratum, we have

$$\bar{b}_{sm} = \frac{1}{N} \left\{ \sum_{i=0}^{N-M-1} M + \sum_{i=N-M}^{N-1} (N - i) \right\}$$

$$= \frac{M}{N} \left\{ N - \frac{M-1}{2} \right\}.$$

Let us now assume that, as described in section 3.2, we select a sample using a design with proportional allocation in which all the strata are of the same size and in which the sizes of H strata are all equal. In such a design, the mean horizon has a value of

$$\bar{b}_{strat} = \frac{1}{2} \left(\frac{N}{H} + 1 \right).$$

A change in the mean horizon does not fundamentally affect the first order inclusion probabilities. The second order inclusion probabilities, on the other hand, are strongly affected by a change of horizon. In fact, it can easily be seen that the smaller the mean horizon, the smaller the probability of selecting two close individuals. (Two individuals are said to be close if the absolute value of the difference of their serial numbers in the data file is small.) Intuitively, we can expect the moving stratum algorithm to have a stratification effect similar to that of a stratified design with proportional allocation having the same mean horizon, *i.e.*, when

$$\bar{b}_{strat} = \bar{b}_{sm},$$

or in other words, when

$$M = N + \frac{1}{2} - \sqrt{\frac{1}{4} + N^2 \frac{H-1}{H}}. \qquad (17)$$

When N is large in relation to M, we have approximately

$$M \approx \frac{2N}{H}.$$

For each series of simulations presented in the Appendix (Table 2), the sizes of the moving strata (case: *sm*) were fixed in terms of the number of strata (case: *strat*) in such a way that the mean horizons of the two designs were identical in terms of expression (17). It is observed that, in such a case, the increased precision (compared to that of the simple design) derived from the moving stratum algorithm is of the same order of magnitude as that derived by means of stratification.

5. COMMENTS

The simulations that were carried out clearly show that the moving stratification algorithm yields a stratification effect of the same type as classical stratification with proportional allocation. This algorithm makes it possible to study the delicate problem of subdividing a continuous variable into strata. The estimators of the mean that are proposed are slightly biased. However, as long as $M \geq 10N/n$, simulations show that it is extremely rare for at least one of the c_i to fall outside of $[0,1]$. Moreover, we have shown that even when that probability is not zero, the bias of the estimator that we propose is negligible as long as $M \geq 3N/n$.

ACKNOWLEDGEMENTS

The author wishes to thank Pierre Lavallée for comments made on previous drafts of this paper. The author also wishes to thank an associate editor and a referee who provided numerous constructive comments that significantly improved this paper.

APPENDIX 1

Demonstration of the Lemmas and Propositions

Demonstration of Lemma 3

$$\begin{aligned}\operatorname{Var}[n_{i+1}] &= \operatorname{Var}[n_i] + \operatorname{Var}[I_{i+1}]\\ &\quad + 2E\left(E\left\{\left(n_i - i\frac{n}{N}\right)E\left[I_{i+1} - \frac{n}{N} \mid n_i\right]\right\}\right).\end{aligned}$$

Since

$$\begin{aligned}&2E\left[E\left\{\left(n_i - i\frac{n}{N}\right)E\left[\left(I_{i+1} - \frac{n}{N}\right) \mid n_i\right]\right\}\right]\\ &\quad = 2E\left[\left(n_i - i\frac{n}{N}\right)\left(\frac{(b_i + i)n/N - n_i}{b_i} - \frac{n}{N}\right)\right]\\ &\quad = \frac{-2}{b_i}\operatorname{Var}[n_i],\end{aligned}$$

we obtain

$$\operatorname{Var}[n_{i+1}] = \operatorname{Var}[n_i]\frac{b_i - 2}{b_i} + \frac{n}{N}\frac{N - n}{N},$$

$$i = 1, \ldots, N - 1. \quad (18)$$

We then show that (3) verifies the recursion equation (18) and the initial condition given by

$$\operatorname{Var}(n_1) = \frac{n}{N}\frac{N - n}{N}.$$

Demonstration of Proposition 1

Case 1: $i = 0$. From lemma 2 we immediately get:

$$\begin{aligned}E[I_k I_1] &= E[E[I_k \mid n_1]\,n_1]\\ &= \frac{n^2}{N^2} - \frac{n}{N}\frac{N - n}{N}\frac{1}{b_{k-1}}\prod_{\ell=1}^{k-2}\frac{b_\ell - 1}{b_\ell}.\end{aligned}$$

Case 2: $i > 0$. Using lemma 2, we obtain:

$$\begin{aligned}&E[I_{i+k}I_{i+1} \mid n_i = t]\\ &\quad = E[I_{i+k} \mid n_{i+1} = t + 1]E[I_{i+1} \mid n_i = t]\\ &\quad = \left\{\frac{n}{N} - \left((t+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}\right\}\\ &\qquad \times \left\{\frac{n}{N} - \left(t - i\frac{n}{N}\right)\frac{1}{b_i}\right\}.\end{aligned}$$

Which means that

$$\begin{aligned}&E[E[I_{i+k}I_{i+1} \mid n_i]]\\ &\quad = E\left\{\frac{n}{N} - \left((n_i+1) - (i+1)\frac{n}{N}\right)\frac{1}{b_{i+k-1}}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}\right\}\\ &\qquad \times \left\{\frac{n}{N} - \left(n_i - i\frac{n}{N}\right)\frac{1}{b_i}\right\}\\ &\quad = \frac{n^2}{N^2} - \frac{1}{b_{i+k-1}}\left\{\frac{n}{N}\frac{N-n}{N} - \frac{\operatorname{Var}[n_i]}{b_i}\right\}\prod_{\ell=i+1}^{i+k-2}\frac{b_\ell - 1}{b_\ell}.\end{aligned}$$

Lemma 3 thus gives us $\operatorname{Var}[n_i]$. We immediately obtain (4).

Demonstration of Proposition 2

Using (13), we have

$$\Pr\left[n - M - \frac{n}{N} < n_{N-M} < \frac{N - n}{N} + n\right] = 1.$$

Therefore,

$$\Pr[0 \le n - n_{N-M} \le M] = 1.$$

Beginning with step $N - M$, the algorithm is a selection-rejection algorithm of the type described in section 3.1. This algorithm yields a sample of exactly $n - n_{N-M}$ observation units during the final M steps. Since $n - n_{N-M} \le M$, this operation raises no difficulty and the algorithm is therefore of fixed size n.

APPENDIX 2

Tables, Bias Upper Bounds and Simulations

Table 1

Value of the Bias Upper Bounds C_α

N	n	Value of the Coefficient C_α				
		$M = \frac{N}{n}$	$M = \frac{2N}{n}$	$M = \frac{3N}{n}$	$M = \frac{4N}{n}$	$M = \frac{5N}{n}$
100	50	0.000000	0.000000	0.000000	0.000000	0.000000
	25	0.057326	0.002610	0.000185	0.000015	0.000001
	12	0.041716	0.002604	0.000235	0.000023	0.000002
	6	0.032227	0.002029	0.000134	0.000005	0.000000
	3	0.023515	0.000645	0.000000		
500	250	0.000000	0.000000	0.000000	0.000000	0.000000
	125	0.129091	0.006002	0.000437	0.000038	0.000004
	62	0.090863	0.005664	0.000534	0.000059	0.000007
	31	0.066891	0.004666	0.000484	0.000059	0.000008
	15	0.048544	0.003586	0.000384	0.000046	0.000006
	7	0.035508	0.002552	0.000215	0.000015	0.000001
	3	0.024046	0.000699	0.000000		
2,500	1,250	0.000000	0.000000	0.000000	0.000000	0.000000
	625	0.289060	0.013495	0.000987	0.000086	0.000008
	312	0.202458	0.012607	0.001190	0.000133	0.000016
	156	0.147113	0.010234	0.001064	0.000130	0.000017
	78	0.105662	0.007742	0.000841	0.000107	0.000015
	39	0.075975	0.005719	0.000634	0.000082	0.000012
	19	0.054525	0.004174	0.000466	0.000060	0.000008
	9	0.039560	0.003014	0.000301	0.000029	0.000002
	4	0.028388	0.001451	0.000034	0.000000	
12,500	3,125	0.646539	0.030208	0.002211	0.000193	0.000018
	1,562	0.452450	0.028177	0.002661	0.000297	0.000036
	781	0.327879	0.022798	0.002371	0.000290	0.000039
	390	0.234114	0.017131	0.001863	0.000238	0.000033
	195	0.166626	0.012500	0.001388	0.000181	0.000026
	97	0.118357	0.008995	0.001009	0.000133	0.000019
	48	0.084217	0.006452	0.000727	0.000096	0.000014
	24	0.060797	0.004689	0.000529	0.000069	0.000010
	12	0.044677	0.003461	0.000377	0.000044	0.000005
	6	0.033727	0.002356	0.000173	0.000008	0.000000
	3	0.024172	0.000712	0.000000		
62,500	3 906	0.732684	0.050942	0.005299	0.000649	0.000087
	1,953	0.522918	0.038250	0.004159	0.000531	0.000074
	976	0.371301	0.027833	0.003092	0.000403	0.000057
	488	0.263300	0.019979	0.002243	0.000295	0.000042
	244	0.186736	0.014259	0.001609	0.000213	0.000031
	122	0.132653	0.010168	0.001150	0.000152	0.000022
	61	0.094601	0.007273	0.000823	0.000109	0.000016
	30	0.067467	0.005207	0.000590	0.000078	0.000011
	15	0.049227	0.003820	0.000427	0.000054	0.000007
	7	0.035847	0.002637	0.000227	0.000016	0.000001
	3	0.024176	0.000713	0.000000		
312,500	4,882	0.829762	0.062191	0.006909	0.000901	0.000128
	2,441	0.587909	0.044596	0.005006	0.000659	0.000095
	1,220	0.416165	0.031758	0.003583	0.000474	0.000068
	610	0.294647	0.022555	0.002551	0.000339	0.000049
	305	0.208743	0.016008	0.001813	0.000241	0.000035
	152	0.147877	0.011356	0.001287	0.000171	0.000025
	76	0.105272	0.008098	0.000918	0.000122	0.000018
	38	0.075422	0.005817	0.000659	0.000087	0.000013
	19	0.054695	0.004238	0.000479	0.000062	0.000009
	9	0.039644	0.003038	0.000305	0.000030	0.000002
	4	0.028427	0.001457	0.000034	0.000000	

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification

ρ^2	Plan	Parameters	$E_{sim}\widehat{\text{Var}}\,\hat{\bar{y}}$	$\text{Var}\hat{\bar{y}}$	$EQM_{sim}\hat{\bar{y}}$
0.0	*sm*	$M = 18.83N/n$	0.01318	0.01317	0.01301
	srs		0.01317	0.01316	0.01296
	strat	$H = 2$	0.01319	0.01319	0.01318
0.2	*sm*	$M = 18.83N/n$	0.01210	0.01210	0.01187
	srs		0.01316	0.01316	0.01287
	strat	$H = 2$	0.01172	0.01188	0.01164
0.4	*sm*	$M = 18.83N/n$	0.01073	0.01073	0.01080
	srs		0.01316	0.01316	0.01320
	strat	$H = 2$	0.00943	0.00929	0.00946
0.6	*sm*	$M = 18.83N/n$	0.00957	0.00957	0.00954
	srs		0.01315	0.01316	0.01301
	strat	$H = 2$	0.00783	0.00778	0.00774
0.8	*sm*	$M = 18.83N/n$	0.00839	0.00839	0.00839
	srs		0.01315	0.01316	0.01322
	strat	$H = 2$	0.00630	0.00624	0.00622
1.0	*sm*	$M = 18.83N/n$	0.00757	0.00757	0.00760
	srs		0.01314	0.01316	0.01319
	strat	$H = 2$	0.00514	0.00508	0.00513
0.0	*sm*	$M = 8.65N/n$	0.01319	0.01319	0.01317
	srs		0.01317	0.01316	0.01296
	strat	$H = 4$	0.01320	0.01318	0.01316
0.2	*sm*	$M = 8.65N/n$	0.01107	0.01107	0.01084
	srs		0.01316	0.01316	0.01287
	strat	$H = 4$	0.01080	0.01076	0.01054
0.4	*sm*	$M = 8.65N/n$	0.00876	0.00876	0.00882
	srs		0.01316	0.01316	0.01320
	strat	$H = 4$	0.00811	0.00793	0.00796
0.6	*sm*	$M = 8.65N/n$	0.00695	0.00694	0.00688
	srs		0.01315	0.01316	0.01301
	strat	$H = 4$	0.00637	0.00639	0.00632
0.8	*sm*	$M = 8.65N/n$	0.00484	0.00484	0.00485
	srs		0.01315	0.01316	0.01322
	strat	$H = 4$	0.00402	0.00391	0.00390
1.0	*sm*	$M = 8.65N/n$	0.00312	0.00312	0.00313
	srs		0.01314	0.01316	0.01319
	strat	$H = 4$	0.00206	0.00197	0.00197
0.0	*sm*	$M = 4.21N/n$	0.01317	0.01317	0.01316
	srs		0.01317	0.01316	0.01296
	strat	$H = 8$	0.01321	0.01324	0.01325
0.2	*sm*	$M = 4.21N/n$	0.01067	0.01067	0.01046
	srs		0.01316	0.01316	0.01287
	strat	$H = 8$	0.01055	0.01047	0.01025
0.4	*sm*	$M = 4.21N/n$	0.00810	0.00809	0.00808
	srs		0.01316	0.01316	0.01320
	strat	$H = 8$	0.00794	0.00789	0.00789
0.6	*sm*	$M = 4.21N/n$	0.00592	0.00592	0.00588
	srs		0.01315	0.01316	0.01301
	strat	$H = 8$	0.00575	0.00564	0.00561
0.8	*sm*	$M = 4.21N/n$	0.00344	0.00344	0.00345
	srs		0.01315	0.01316	0.01322
	strat	$H = 8$	0.00315	0.00311	0.00308
1.0	*sm*	$M = 4.21N/n$	0.00124	0.00124	0.00125
	srs		0.01314	0.01316	0.01319
	strat	$H = 8$	0.00085	0.00079	0.00080

Table 2

Results of the Simulations, Simple Design, Stratification and Moving Stratification - end

ρ^2	Plan	Parameters	$E_{sim}\widehat{\mathrm{Var}}\,\hat{\bar{y}}$	$\mathrm{Var}\,\hat{\bar{y}}$	$EQM_{sim}\hat{\bar{y}}$
0.0	*sm*	$M = 2.11N/n$	0.01319	0.01319	0.01328
	srs		0.01315	0.01316	0.01332
	strat	$H = 16$	0.01315	0.01308	0.01331
0.2	*sm*	$M = 2.11N/n$	0.01038	0.01036	0.01021
	srs		0.01317	0.01316	0.01334
	strat	$H = 16$	0.01034	0.01034	0.01025
0.4	*sm*	$M = 2.11N/n$	0.00796	0.00796	0.00792
	srs		0.01316	0.01316	0.01323
	strat	$H = 16$	0.00790	0.00801	0.00794
0.6	*sm*	$M = 2.11N/n$	0.00572	0.00573	0.00561
	srs		0.01315	0.01316	0.01299
	strat	$H = 16$	0.00568	0.00572	0.00563
0.8	*sm*	$M = 2.11N/n$	0.00295	0.00294	0.00290
	srs		0.01317	0.01316	0.01325
	strat	$H = 16$	0.00287	0.00288	0.00285
1.0	*sm*	$M = 2.11N/n$	0.00048	0.00048	0.00048
	srs		0.01317	0.01316	0.01335
	strat	$H = 16$	0.00037	0.00034	0.00034
0.0	*sm*	$M = 1.09N/n$	0.01325	0.01316	0.01310
	srs		0.01313	0.01316	0.01317
	strat	$H = 32$	0.01201	0.01239	0.01302
0.2	*sm*	$M = 1.09N/n$	0.01070	0.01062	0.01064
	srs		0.01313	0.01316	0.01316
	strat	$H = 32$	0.00972	0.01018	0.01083
0.4	*sm*	$M = 1.09N/n$	0.00807	0.00803	0.00811
	srs		0.01315	0.01316	0.01309
	strat	$H = 32$	0.00732	0.00751	0.00803
0.6	*sm*	$M = 1.09N/n$	0.00538	0.00534	0.00536
	srs		0.01315	0.01316	0.01310
	strat	$H = 32$	0.00484	0.00484	0.00543
0.8	*sm*	$M = 1.09N/n$	0.00283	0.00281	0.00276
	srs		0.01317	0.01316	0.01283
	strat	$H = 32$	0.00255	0.00276	0.00280
1.0	*sm*	$M = 1.09N/n$	0.00016	0.00016	0.00017
	srs		0.01317	0.01316	0.01304
	strat	$H = 32$	0.00012	0.00007	0.00011

REFERENCES

BEBBINGTON, A.C. (1975). A simple method of drawing a sample without replacement. *Applied Statistics*, 24, 136.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: Wiley.

DEVILLE, J.-C., and GROSBRAS, J.-M. (1987). Algorithmes de tirage. In *Les sondages*. Droesbeke, J.-J., Fichet, B., and Tassi, P. (Eds.). Paris: Economica, 209-233.

FAN, C.T., MULLER, M.E., and REZUCHA, I. (1962). Development of sampling plans by using sequential (item by item) selection techniques and digital computers. *Journal of the American Statistical Association*, 57, 387-402.

HÁJEK, J. (1971). Comment on an essay of D. Basu. In *Foundations of Statistical Inference*. Godambe V.P., and Sprott, D.A. (Eds). Toronto: Holt, Rinehart and Winston.

McLEOD, A.I., and BELLHOUSE, D.R. (1983). A convenient algorithm for drawing a simple random sampling. *Applied Statistics*, 32, 182-184.

SUNTER, A.B. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

SUNTER, A.B. (1986). Solutions to the problem of unequal probability sampling without replacement. *International Statistical Revue*, 54, 33-50.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, B, 15, 235-261.