Bernoulli **18**(4), 2012, 1320–1340
DOI: 10.3150/11-BEJ380

On a characterization of ordered pivotal sampling

GUILLAUME CHAUVET

Crest (Ensai), Campus de Ker Lann, 35170 Bruz, France. E-mail: chauvet@ensai.fr

When auxiliary information is available at the design stage, samples may be selected by means of balanced sampling. Deville and Tillé proposed in 2004 a general algorithm to perform balanced sampling, named the cube method. In this paper, we are interested in a particular case of the cube method named pivotal sampling, and first described by Deville and Tillé in 1998. We show that this sampling algorithm, when applied to units ranked in a fixed order, is equivalent to Deville's systematic sampling, in the sense that both algorithms lead to the same sampling design. This characterization enables the computation of the second-order inclusion probabilities for pivotal sampling. We show that the pivotal sampling enables to take account of an appropriate ordering of the units to achieve a variance reduction, while limiting the loss of efficiency if the ordering is not appropriate.

Keywords: balanced sampling; cube method; design effect; sampling algorithm; second order inclusion probabilities; unequal probabilities

1. Introduction

When auxiliary information is available at the design stage, samples may be selected by means of balanced sampling. The variance of the Horvitz–Thompson (HT) estimator is then reduced, since it is approximately given by that of the residuals of the variable of interest on the balancing variables. Deville and Tillé [6] proposed a general algorithm for balanced sampling, named the *cube method*. This sampling algorithm enables the selection of balanced samples with any number of balancing variables, and any prescribed set of inclusion probabilities.

In order to measure the gain in efficiency provided by the cube method, Deville and Tillé [7] proposed several variance approximations. They suppose that the sampling design is exactly balanced, and performed with maximum entropy among sampling designs balanced on the same balancing variables, with the same inclusion probabilities. Then, under an additional assumption of asymptotic normality of the multivariate HT-estimator under Poisson sampling, the variance approximations are derived. The assumption of exact balancing may be closely respected, if the number of balancing variables remains small with regard to the sample size; otherwise, the balancing error must be taken into account in variance estimation, see Breidt and Chauvet [9]. The second assumption is related to the entropy of the sampling design: the variance approximations proposed by Deville and Tillé [7] are unlikely to hold if this assumption is not satisfied.

A practical way to increase the entropy of a sampling design is to sort the population randomly before sampling. However, this preliminary randomization step is not systematically included in the sampling process. This is a common practice to sort the population with respect to some auxiliary variable before the sampling, so as to benefit from a stratification effect. In France,

1350-7265 © 2012 ISI/BS

Census surveys are conducted annually; the detailed methodology is described in Godinot [8]. Each large municipality (10 000 inhabitants or more in 1999) is the subject of an independent sampling design and is stratified according to the type of address (large addresses, new addresses, or other addresses). In each stratum, the addresses are divided into 5 rotation groups. Each year, all the addresses within one rotation group (for the strata of large addresses and new addresses) or within a sub-sample (for the stratum of other addresses) are surveyed. In the stratum of other addresses, the sub-sample is obtained by first, sorting the addresses with respect to the descending number of dwellings, and then, applying the cube method. In such cases, the conditions for the variance approximations proposed by Deville and Tillé [7] to hold are clearly not respected.

We are interested in a particular case of the cube method, called *pivotal sampling* (Deville and Tillé [5]), obtained when the only balancing condition is given by the variable of inclusion probabilities. That is, the cube method with the sole fixed-size constraint amounts to pivotal sampling. This algorithm is an exact sampling procedure, which respects a prescribed set of inclusion probabilities, is strictly without replacement and leads to fixed-size designs. In this paper, we show that the pivotal sampling algorithm, when applied to units ranked in a fixed order, is equivalent to an algorithm proposed in Deville [4], and known in the literature as *Deville's systematic sampling* (Tillé [13]). The two algorithms are equivalent, in the sense that both lead to the same sampling design. In particular, the computation of the second-order inclusion probabilities developed in Deville [4] may be readily applied to pivotal sampling. This provides an answer to a problem raised by Bondesson and Grafström [1], page 7. Deville's systematic sampling has similarities with Markov chain designs introduced by Breidt [2]. It has found uses in the context of longitudinal surveys, see Nedyalkova, Qualité, and Tillé [11].

The paper is organized as follows. In Section 2, the notation is defined. Ordered pivotal sampling and Deville's systematic sampling are presented in Sections 3 and 4, respectively, and some useful results are derived. The second-order inclusion probabilities for ordered pivotal sampling are given in Section 5. Some results which illustrate the practical interest of ordered pivotal sampling are presented in Section 6.

2. Notation

Consider a finite population U consisting of N sampling units that may be represented by integers $k = 1, \dots, N$. We assume that the order of the units in the population is fixed prior to sampling, and may be confounded with the natural order of their indexes. A sample s, defined as a subset of U, is selected with inclusion probabilities $\pi = (\pi_1, \dots, \pi_N)'$. We assume without loss of generality that $0 < \pi_k < 1$ for any unit k in U, with $n = \sum_{k \in U} \pi_k$ the sample size. Let π_{kl} denote the probability that units k and l are selected jointly in the sample.

We define $V_k = \sum_{l=1}^{k} \pi_l$ for any unit $k \in U$, with $V_0 = 0$. A unit k is said to be *cross-border* if $V_{k-1} \leq i$ and $V_k > i$ for some nonnegative integer i. The cross-border units are denoted as k_i, $i = 1, \dots, n-1$, and we note $a_i = i - V_{k_i - 1}$ and $b_i = V_{k_i} - i$. The microstratum U_i, $i = 1, \dots, n$, is defined as

$$U_i = \{k \in U; k_{i-1} \leq k \leq k_i\}, \tag{2.1}$$

with $k_0 = 0$ and $k_n = N + 1$. To fix ideas, useful quantities for population U are presented in Figure 1.

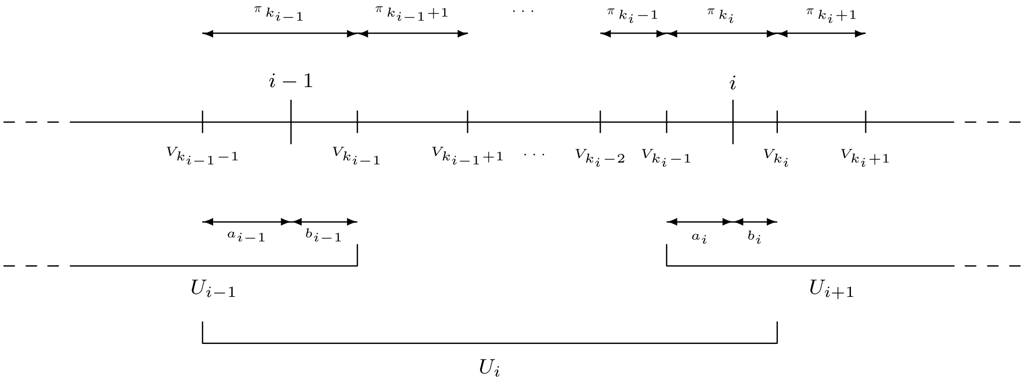

Figure 1. Inclusion probabilities and cross-border units in microstratum U_i, for population U.

The microstrata are generally overlapping, since one cross-border unit may belong to two adjacent microstrata: the cross-border unit k_i belongs both to the microstratum U_i (with an associated probability a_i) and to the microstratum U_{i+1} (with an associated probability b_i). In the particular case when $V_{k_i} = i$, we have $b_i = 0$. To avoid the introduction of specific notations for such cases, we consider in Sections 2–5 that, in this situation, the cross-border unit k_i belongs to the microstratum U_{i+1} as a "phantom unit," that is, with an associated probability equal to 0. In Section 6, we simply consider that the cross-border unit k_i belongs to the microstratum U_i only in such situations.

The N sampling units are grouped to obtain a population $U_c = \{u_1, \ldots, u_{2n-1}\}$ of clusters. There are the clusters of cross-border units ($n-1$ singletons), denoted as u_{2i} with associated probability $\phi_{2i} = \pi_{k_i}$ for $i = 1, \ldots, n-1$. There are the n clusters of units that are not cross-borders and that are between two consecutive integers, denoted as u_{2i-1} with associated probability $\phi_{2i-1} = V_{k_i-1} - V_{k_{i-1}}$, for $i = 1, \ldots, n$. We note $\psi = (\phi_1, \ldots, \phi_{2n-1})'$. To fix ideas, useful quantities for population U_c are presented in Figure 2. If (at least) one of the cross-border units

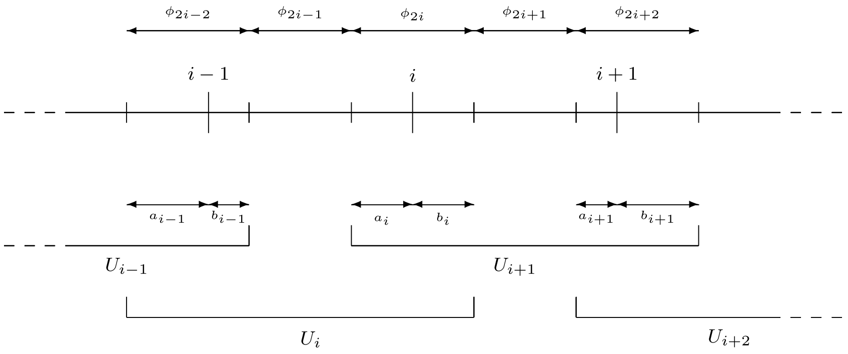

Figure 2. Inclusion probabilities and cross-border units in microstrata U_i and U_{i+1} for population U_c.

Table 1. Clusters and associated probabilities for a population of size 8

i	1	2	3	4	5	6	7
u_i	$\{1,2\}$	$\{3\}$	$\{4\}$	$\{5\}$	$\emptyset$	$\{6\}$	$\{7,8\}$
ϕ_i	0.7	0.3	0.4	0.9	0	0.8	0.9

in U_i has a large inclusion probability, there may not exist any non-cross-border unit between integers $i-1$ and i, so that the cluster u_{2i-1} is empty. To avoid the need for specific notations for such cases, we may view this situation as a particular case of our framework by allowing a cluster u_{2i-1} to be a "phantom cluster," that is, an empty cluster with associated probability ϕ_{2i-1} equal to 0. For example, suppose that $N=8$, $n=4$ and $\pi=(0.2,0.5,0.3,0.4,0.9,0.8,0.5,0.4)'$. We obtain the 4 microstrata $U_1=\{1,2,3\}$, $U_2=\{3,4,5\}$, $U_3=\{5,6\}$ and $U_4=\{6,7,8\}$. In particular, we have $a_1=0.3=\pi_3$ and $b_1=0$, so that the cross-border unit 3 is a phantom unit for the microstratum U_2. Also, we obtain 7 clusters (see Table 1): the cluster u_5 is empty, with an associated probability equal to zero.

3. Ordered pivotal sampling

A general algorithm for pivotal sampling is described in Deville and Tillé [5]. In the version presented in Algorithm 1, the order of the sampling units is explicitly taken into account. We call it *ordered pivotal sampling* to avoid confusion. At each step, one or two coordinates of $\pi(t)$ are randomly rounded to 0 or 1, and remain there forever. In at most N steps, the final sample is obtained.

Roughly speaking, the algorithm may be summarized as follows. At the beginning, in microstratum U_1 ($i=1$), the two first units 1 and 2 fight, the loser is definitely eliminated while the survivor (denoted as J_0) gets the sum of their probabilities and then faces the following unit. The fights go on until the accumulated probability exceeds 1, which occurs at time $t=k_1$ when the survivor J_0 faces the cross-border unit k_1. One of the two remaining units, denoted as W_1, wins and is then definitely selected in the sample while the other one, denoted as J_1, jumps to the microstratum U_2.

More generally, in microstratum U_i, the first unit k_{i-1} is replaced with the unit J_{i-1} which jumps from the microstratum U_{i-1}. The units J_{i-1} and $k_{i-1}+1$ fight, the survivor gets the sum of their probabilities and then faces the next unit. The fights go on until the survivor J_{i-1} faces the cross-border unit k_i. One of the two remaining units (W_i) wins and is then definitely selected in the sample while the other one (J_i) jumps to the following microstratum. Lemma 3.1 states that Algorithm 1 may alternatively be seen as a two-stage procedure. The proof follows from definition, and is thus omitted.

Lemma 3.1. *Ordered pivotal sampling with parameter π may be obtained by two-stage sampling, where a sample s_c of n clusters is first selected in U_c by means of ordered pivotal sampling with parameter ψ, and one unit k is then selected in each $u_j \in s_c$ with a probability proportional to π_k.*

Algorithm 1 Ordered Pivotal Sampling with parameter $\boldsymbol{\pi}$

```
1. We initialize with $i = 1$, $J_0 = 1$ and $\pi(0) = \pi$.
2. For $t = 2, \dots, N$, do:
   (a) If $m \in U \setminus \{J_{i-1}, t\}$, then $\pi_m(t) = \pi_m(t-1)$.
   (b) If $\pi_{J_{i-1}}(t-1) + \pi_t(t-1) < 1$, let $\lambda_1(t) = \frac{\pi_{J_{i-1}}(t-1)}{\pi_{J_{i-1}}(t-1)+\pi_t(t-1)}$. Then
       i. with probability $\lambda_1(t)$, let
          $$[\pi_{J_{i-1}}(t), \pi_t(t)] = [\pi_{J_{i-1}}(t-1) + \pi_t(t-1), 0];$$
       ii. with probability $1 - \lambda_1(t)$, let $J_{i-1} = t$ and
          $$[\pi_{J_{i-1}}(t), \pi_t(t)] = [0, \pi_{J_{i-1}}(t-1) + \pi_t(t-1)].$$
   (c) If $\pi_{J_{i-1}}(t-1) + \pi_t(t-1) \geq 1$, let $\lambda_1(t) = \frac{1-\pi_t(t-1)}{2-\pi_{J_{i-1}}(t-1)-\pi_t(t-1)}$. Then
       i. with probability $\lambda_1(t)$, let $W_i = J_{i-1}$, let $J_i = t$ and
          $$[\pi_{J_{i-1}}(t), \pi_t(t)] = [1, \pi_{J_{i-1}}(t-1) + \pi_t(t-1) - 1];$$
       ii. with probability $1 - \lambda_1(t)$, let $W_i = t$, let $J_i = J_{i-1}$ and
          $$[\pi_{J_{i-1}}(t), \pi_t(t)] = [\pi_{J_{i-1}}(t-1) + \pi_t(t-1) - 1, 1];$$
       iii. let $i = i + 1$.
3. The sample is given by $\{W_1, \dots, W_n\}$.
```

We assume that a sample S_{op} is selected in U_c by means of ordered pivotal sampling with parameter ψ, and we let $X_1 < \cdots < X_n$ denote the units selected in the sample, ranked in ascending order. Lemma 3.2 states useful relations between on the one hand, the sampled units X_i, and on the other hand, the winners W_i and jumpers J_i. Lemma 3.3 gives the probabilities for the different outcomes in the case of a non-cross-border unit u_{2i-1}.

Lemma 3.2. *In case of ordered pivotal sampling with parameter* ψ, *we have*

$$\{X_i = u_{2i-2}\} \Rightarrow \big\{J_{i-1} \in \{X_1, \dots, X_i\}\big\}, \tag{3.1}$$

$$\{X_i = u_{2i-1}\} \Rightarrow \{W_i = u_{2i-1}\} \cup \{J_i = u_{2i-1}\}, \tag{3.2}$$

$$\{X_i = u_{2i}\} \Rightarrow \big\{J_i \notin \{X_1, \dots, X_i\}\big\}. \tag{3.3}$$

Proof. Assume that $X_i = u_{2i-2}$. This implies that i units exactly are selected in the $i-1$ first microstrata $U_1, \dots, U_{i-1}$. On the other hand, if $J_{i-1} \notin \{X_1, \dots, X_i\}$ the unit J_{i-1} is not selected in the sample so that at most $i-1$ units are selected in $U_1, \dots, U_{i-1}$. This proves (3.1), and by a similar argument we obtain (3.3). It is easily seen that (3.2) holds, since the selection of u_{2i-1} implies that this unit is either the winner W_i or the jumper J_i in the microstratum U_i. □

Lemma 3.3. *In case of ordered pivotal sampling with parameter* ψ, *we have*

$$pr(W_i = u_{2i-1}) = \frac{(1 - a_i - b_{i-1})(1 - a_i - b_i)}{(1 - a_i)(1 - b_i)}, \tag{3.4}$$

$$pr(J_i = u_{2i-1}) = \frac{a_i(1 - a_i - b_{i-1})}{(1 - a_i)(1 - b_i)}, \tag{3.5}$$

$$pr(X_i = u_{2i-1}) = 1 - a_i - b_{i-1}. \tag{3.6}$$

Proof. The event

$$\{W_i = u_{2i-1}\}$$

may be alternatively interpreted as follows: in the fight between J_{i-1} and u_{2i-1}, the unit u_{2i-1} survives; then in the next fight, the unit u_{2i-1} is the selected unit W_i, while the unit u_{2i} is the jumping unit J_i. Consequently, we have:

$$pr(W_i = u_{2i-1}) = \frac{1 - b_{i-1} - a_i}{1 - a_i} \times \frac{1 - a_i - b_i}{1 - b_i},$$

which gives (3.4). Similarly, we obtain

$$pr(J_i = u_{2i-1}) = \frac{1 - b_{i-1} - a_i}{1 - a_i} \times \frac{a_i}{1 - b_i},$$

which gives (3.5). We now consider equation (3.6). Since

$$\{X_i = u_{2i-1}\} \Rightarrow \{u_{2i-1} \in S_{op}\}$$

and

$$pr(u_{2i-1} \in S_{op}) = 1 - a_i - b_{i-1},$$

it suffices to show that

$$\{u_{2i-1} \in S_{op}\} \Rightarrow \{X_i = u_{2i-1}\}. \tag{3.7}$$

Since $\{u_{2i-1} \in S_{op}\}$ implies that u_{2i-1} survives in its duel against J_{i-1}, this in turn implies that $J_{i-1} \notin \{X_1, \dots, X_i\}$. In other words, $\{u_{2i-1} \in S_{op}\}$ implies that exactly $i - 1$ units smaller than u_{2i-1} were selected, which proves (3.7). □

Finally, let $U_{c,i} = \{u_{2i-2}, \dots, u_{2n-1}\}$, $\psi_i = (b_{i-1}, \phi_{2i-1}, \dots, \phi_j, \dots, \phi_{2n-1})'$, and $S_{op,i}$ be a random sample selected in $U_{c,i}$ by means of ordered pivotal sampling with parameter ψ_i. Lemma 3.4 establishes some relations for conditional inclusion probabilities in $S_{op,i}$ of the first units in $U_{c,i}$.

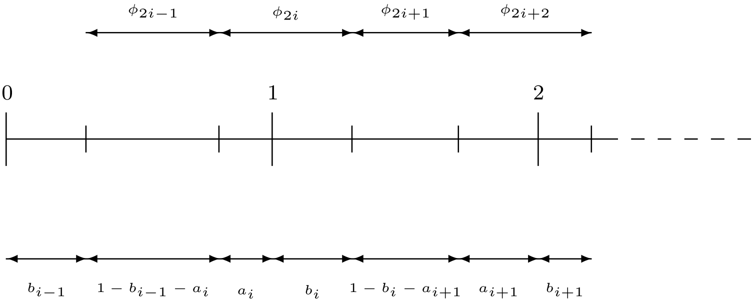

Figure 3. Inclusion probabilities and cross-border units in the two first microstrata of population $U_{c,i}$.

Lemma 3.4.

$$pr(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i}) = \frac{b_i}{1-a_i}, \tag{3.8}$$

$$pr(u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i}) = \frac{(1-a_i-b_i)(1-b_i-a_{i+1})}{(1-a_i)(1-b_i)}, \tag{3.9}$$

$$pr(u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i}) = \frac{(1-a_i-b_i)a_{i+1}}{(1-a_i)(1-b_i)}. \tag{3.10}$$

Proof. To fix ideas, the first units in population $U_{c,i}$ and related quantities are presented in Figure 3.

We first consider equation (3.8). Since b_{i-1} is the first-order inclusion probability of unit u_{2i-2} in sample $S_{op,i}$, we have

$$pr(u_{2i-2} \in S_{op,i}) = b_{i-1}. \tag{3.11}$$

On the other hand, the event

$$\{u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\}$$

may be alternatively interpreted as follows: in the first fight, the unit u_{2i-2} survives against the unit u_{2i-1}; in the second fight, any of the two units u_{2i-2} or u_{2i} is the selected unit W_1, while the other is the jumping unit J_1; then, the jumping unit J_1 is selected during one of the following fights. Consequently, we have:

$$pr(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}) = \frac{b_{i-1}}{1-a_i} \times 1 \times b_i, \tag{3.12}$$

and equation (3.8) follows from (3.11) and (3.12). We now consider equation (3.9). The event

$$\{u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\}$$

may be interpreted as follows: in the first fight, the unit u_{2i-2} survives against the unit u_{2i-1}; in the second fight, u_{2i-2} is the selected unit W_1, while u_{2i} is the jumping unit J_1; in the third fight, the unit u_{2i+1} survives against the unit u_{2i}; then, the unit u_{2i+1} is selected during one of the following fights. Consequently, we have:

$$\begin{aligned} &pr(u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}) \\ &\quad = \frac{b_{i-1}}{1-a_i} \times \frac{1-a_i-b_i}{1-b_i} \times \frac{1-b_i-a_{i+1}}{1-a_{i+1}} \times (1-a_{i+1}) \\ &\quad = \frac{b_{i-1}(1-a_i-b_i)(1-b_i-a_{i+1})}{(1-a_i)(1-b_i)}, \end{aligned} \tag{3.13}$$

which, together with (3.11), leads to (3.9). Finally, we consider equation (3.10). The event

$$\{u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\}$$

may be interpreted as follows: in the first fight, the unit u_{2i-2} survives against the unit u_{2i-1}; in the second fight, u_{2i-2} is the selected unit W_1, while u_{2i} is the jumping unit J_1; in the third fight, any of the two units $J_i = u_{2i}$ or u_{2i+1} survives; in the fourth fight, u_{2i+2} is the selected unit W_2, while the other unit is the jumper J_2; then, the unit J_2 is not selected during one of the following fights. Consequently, we have:

$$\begin{aligned} &pr(u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}) \\ &\quad = \frac{b_{i-1}}{1-a_i} \times \frac{1-a_i-b_i}{1-b_i} \times 1 \times \frac{a_{i+1}}{1-b_{i+1}} \times (1-b_{i+1}) \\ &\quad = \frac{b_{i-1}(1-a_i-b_i)a_{i+1}}{(1-a_i)(1-b_i)}, \end{aligned} \tag{3.14}$$

which gives (3.10). □

4. Deville's systematic sampling

The sampling algorithm known in the literature as Deville's systematic sampling (Deville [4]; Tillé [13]) is presented in Algorithm 2. This algorithm proceeds in n sub-samplings of size 1 in the microstrata $U_1, \ldots, U_n$, and the random variables w_i which indicate the sampled units are generated so that a cross-border unit k_{i-1} may not be selected twice in the sample: at step i, one unit denoted as Y_i is drawn in U_i if k_{i-1} was not selected at step $i-1$, and in $U_i \setminus \{k_{i-1}\}$ otherwise. This sampling algorithm may be particularly useful in the context of business surveys, when a fine stratification is used leading to small and possibly non-integer sample size inside

Algorithm 2 Deville's systematic sampling with parameter $\boldsymbol{\pi}$

```
At step 1:
1. A distributed Uniform(0, 1) random variable $w_1$ is generated.
2. The unit $k$ is selected if $V_{k-1} \le w_1 < V_k$.

At step $i$:
1. A random variable $w_i$ is generated:
    (a) if unit $k_{i-1}$ was selected at step $i-1$, then $w_i$ is generated according to a distributed Uniform$(b_{i-1}, 1)$ random variable,
    (b) otherwise, $w_i$ is generated:
        • according to a distributed Uniform$(0, b_{i-1})$ random variable with probability $a_{i-1}b_{i-1}\{(1-a_{i-1})(1-b_{i-1})\}^{-1}$,
        • according to a distributed Uniform(0, 1) random variable with probability $1 - a_{i-1}b_{i-1}\{(1-a_{i-1})(1-b_{i-1})\}^{-1}$.
2. The unit $k$ is selected if $V_{k-1} \le w_i + (i-1) < V_k$.
```

(micro)strata. Deville's systematic sampling directly handles the rounding problem, since any unit for which the sampling outcome is still undecided is moved to the next stratum, where the final sampling decision is then obtained. Lemma 4.1 follows from the definition of Algorithm 2.

Lemma 4.1. *Deville's systematic sampling with parameter π may be obtained by two-stage sampling, where a sample s_c of n clusters is first selected in U_c by means of Deville's systematic sampling with parameter ψ, and one unit k is then selected in each $u_j \in s_c$ with a probability proportional to π_k.*

Assume that a sample is selected in U_c by means of Deville's systematic sampling with parameter ψ. The random variable Y_{i+1} which gives the result of the sampling in the microstratum U_{i+1} only depends on the outcome of step i, so that

$$pr(Y_{i+1} = u_j | Y_1, \dots, Y_i) = pr(Y_{i+1} = u_j | Y_i). \tag{4.1}$$

The different cases for the transition probabilities in (4.1) easily follow from the definition of Algorithm 2, and are given below:

$$\begin{aligned} &pr(Y_{i+1} = u_j | Y_1, \dots, Y_{i-1}, Y_i = u_{2i-2}) \\ &= \begin{cases} \dfrac{b_i}{1-a_i}, & j = 2i, \\ \dfrac{(1-b_i-a_{i+1})(1-a_i-b_i)}{(1-a_i)(1-b_i)}, & j = 2i+1, \\ \dfrac{a_{i+1}(1-a_i-b_i)}{(1-a_i)(1-b_i)}, & j = 2i+2, \end{cases} \end{aligned} \tag{4.2}$$

$$pr(Y_{i+1} = u_j | Y_1, \dots, Y_{i-1}, Y_i = u_{2i-1}) \tag{4.3}$$
$$= \begin{cases} \dfrac{b_i}{1-a_i}, & j = 2i, \\ \dfrac{(1-b_i-a_{i+1})(1-a_i-b_i)}{(1-a_i)(1-b_i)}, & j = 2i+1, \\ \dfrac{a_{i+1}(1-a_i-b_i)}{(1-a_i)(1-b_i)}, & j = 2i+2, \end{cases}$$

$$pr(Y_{i+1} = u_j | Y_1, \dots, Y_{i-1}, Y_i = u_{2i}) \tag{4.4}$$
$$= \begin{cases} \dfrac{(1-b_i-a_{i+1})}{(1-b_i)}, & j = 2i+1, \\ \dfrac{a_{i+1}}{(1-b_i)}, & j = 2i+2. \end{cases}$$

5. Second-order inclusion probabilities

We can now formulate our main result.

Theorem 5.1. *Ordered pivotal sampling and Deville's systematic sampling with the same parameter π induce the same sampling design.*

Proof. From Lemmas 3.1 and 4.1, it is sufficient to prove the result in case of ordered systematic sampling and Deville's systematic sampling with parameter ψ in the population U_c. We only need to show that equations (4.2)–(4.4) hold in case of ordered pivotal sampling. Recall that we note

$$U_{c,i} = \{u_{2i-2}, \dots, u_{2n-1}\},$$
$$\psi_i = (b_{i-1}, \phi_{2i-1}, \dots, \phi_j, \dots, \phi_{2n-1})',$$

and that $S_{op,i}$ denotes a random sample selected in $U_{c,i}$ by means of ordered pivotal sampling with parameter ψ_i (see Section 3).

We first consider equation (4.2). From (3.1), we obtain:

$$pr(X_{i+1} = u_{2i} | X_1, \dots, X_{i-1}, X_i = u_{2i-2})$$
$$= pr(X_{i+1} = u_{2i} | X_1, \dots, X_{i-1}, X_i = u_{2i-2}, J_{i-1} \in \{X_1, \dots, X_i\}),$$

which is equivalent to $pr(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i})$, so that the result follows from equation (3.8).

Similarly, we obtain

$$pr(X_{i+1} = u_{2i+1} | X_1, \dots, X_{i-1}, X_i = u_{2i-2})$$
$$= pr(X_{i+1} = u_{2i+1} | X_1, \dots, X_{i-1}, X_i = u_{2i-2}, J_{i-1} \in \{X_1, \dots, X_i\})$$

$$
\begin{aligned}
&\equiv pr(u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i}) \\
&= \frac{(1-a_i-b_i)(1-b_i-a_{i+1})}{(1-a_i)(1-b_i)},
\end{aligned}
$$

where the last line follows from (3.9), and

$$
\begin{aligned}
&pr(X_{i+1} = u_{2i+2} | X_1, \dots, X_{i-1}, X_i = u_{2i-2}) \\
&\quad = pr(X_{i+1} = u_{2i+2} | X_1, \dots, X_{i-1}, X_i = u_{2i-2}, J_{i-1} \in \{X_1, \dots, X_i\}) \\
&\quad \equiv pr(u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} | u_{2i-2} \in S_{op,i}) \\
&\quad = \frac{(1-a_i-b_i)a_{i+1}}{(1-a_i)(1-b_i)},
\end{aligned}
$$

where the last line follows from (3.10). This proves equation (4.2). The proof for equation (4.4) is similar, and is thus omitted.

We now turn to equation (4.3). We introduce some further notation. Let

$$
\begin{aligned}
U_{c,i+1} &= \{u_{2i}, \dots, u_{2n-1}\}, \\
\psi_{i+1} &= (b_i, \phi_{2i+1}, \dots, \phi_j, \dots, \phi_{2n-1})',
\end{aligned}
$$

and let $S_{op,i+1}$ be a random sample selected in $U_{c,i+1}$ by means of ordered pivotal sampling with parameter ψ_{i+1}. We have

$$
\begin{aligned}
&pr(X_{i+1} = u_{2i} | X_1, \dots, X_{i-1}, X_i = u_{2i-1}, W_i = u_{2i-1}) \\
&\quad = pr(X_{i+1} = u_{2i} | X_1, \dots, X_{i-1}, X_i = u_{2i-1}, J_i = u_{2i}) \qquad (5.1) \\
&\quad \equiv pr(u_{2i} \in S_{op,i+1}) = b_i,
\end{aligned}
$$

where the second line in (5.1) comes from

$$\{X_i = u_{2i-1}, W_i = u_{2i-1}\} \Leftrightarrow \{X_i = u_{2i-1}, J_i = u_{2i}\}.$$

Also,

$$pr(X_{i+1} = u_{2i} | X_1, \dots, X_{i-1}, X_i = u_{2i-1}, J_i = u_{2i-1}) = 1, \qquad (5.2)$$

since

$$\{X_i = u_{2i-1}, J_i = u_{2i-1}\} \Rightarrow \{X_i = u_{2i-1}, W_i = u_{2i}\} \Rightarrow \{X_{i+1} = u_{2i}\}.$$

Further,

$$
\begin{aligned}
&pr(W_i = u_{2i-1} | X_1, \dots, X_{i-1}, X_i = u_{2i-1}) \\
&\quad = pr(W_i = u_{2i-1} | X_i = u_{2i-1}) \qquad (5.3)
\end{aligned}
$$

$$
\begin{aligned}
&= pr(X_i = u_{2i-1} | W_i = u_{2i-1}) \frac{pr(W_i = u_{2i-1})}{pr(X_i = u_{2i-1})} \\
&= 1 \times \frac{(1-a_i-b_{i-1})(1-a_i-b_i)\{(1-a_i)(1-b_i)\}^{-1}}{1-a_i-b_{i-1}} \\
&= \frac{1-a_i-b_i}{(1-a_i)(1-b_i)},
\end{aligned}
$$

the fourth line in (5.3) being a consequence of Lemma 3.3. The same reasoning leads to

$$
\begin{aligned}
&pr(J_i = u_{2i-1} | X_1, \ldots, X_{i-1}, X_i = u_{2i-1}) \\
&\quad = pr(J_i = u_{2i-1} | X_i = u_{2i-1}) \\
&\quad = pr(X_i = u_{2i-1} | J_i = u_{2i-1}) \frac{pr(J_i = u_{2i-1})}{pr(X_i = u_{2i-1})} \\
&\quad = b_i \times \frac{a_i(1-a_i-b_{i-1})\{(1-a_i)(1-b_i)\}^{-1}}{1-a_i-b_{i-1}} \\
&\quad = \frac{a_i b_i}{(1-a_i)(1-b_i)}.
\end{aligned}
\tag{5.4}
$$

From equations (5.1)–(5.4), we obtain that

$$
\begin{aligned}
&pr(X_{i+1} = u_{2i} | X_1, \ldots, X_{i-1}, X_i = u_{2i-1}) \\
&\quad = b_i \times \frac{1-a_i-b_i}{(1-a_i)(1-b_i)} + 1 \times \frac{a_i b_i}{(1-a_i)(1-b_i)} \\
&\quad = \frac{b_i}{1-a_i}.
\end{aligned}
$$

Similar computations lead to

$$
pr(X_{i+1} = u_{2i+1} | X_1, \ldots, X_{i-1}, X_i = u_{2i-1}) = \frac{(1-b_i-a_{i+1})(1-a_i-b_i)}{(1-a_i)(1-b_i)}
$$

and

$$
pr(X_{i+1} = u_{2i+2} | X_1, \ldots, X_{i-1}, X_i = u_{2i-1}) = \frac{a_{i+1}(1-a_i-b_i)}{(1-a_i)(1-b_i)},
$$

which proves (4.3). □

Theorem 5.1 implies that ordered pivotal sampling shares the same second-order inclusion probabilities as Deville's systematic sampling. The computation of these probabilities is developed in Deville [4], and is reminded below.

Theorem 5.2 (Deville [4]). *Let k and l be two distinct units in U. If k and l are two non-cross-border units that belong to the same microstratum U_i, then*

$$\pi_{kl} = 0,$$

if k and l are two non-cross-border units that belong to distinct microstrata U_i and U_j, respectively, where $i < j$, then

$$\pi_{kl} = \pi_k \pi_l \{1 - c(i, j)\},$$

if $k = k_{i-1}$ and l is a non-cross-border unit that belongs to the microstratum U_j where $i \leq j$, then

$$\pi_{kl} = \pi_k \pi_l [1 - b_{i-1}(1 - \pi_k)\{\pi_k(1 - b_{i-1})\}^{-1} c(i, j)],$$

if $l = k_{j-1}$ and k is a non-cross-border unit that belongs to the microstratum U_i where $i < j$, then

$$\pi_{kl} = \pi_k \pi_l \{1 - (1 - \pi_l)(1 - b_{j-1})(\pi_l b_{j-1})^{-1} c(i, j)\},$$

if $k = p_{i-1}$ and $l = p_{j-1}$, where $i < j$, then

$$\pi_{kl} = \pi_k \pi_l [1 - b_{i-1}(1 - b_{j-1})(1 - \pi_k)(1 - \pi_l)\{\pi_k \pi_l b_{j-1}(1 - b_{i-1})\}^{-1} c(i, j)],$$

where $c(i, j) = \prod_{l=i}^{j-1} c_l$, $c_l = a_l b_l \{(1 - a_l)(1 - b_l)\}^{-1}$ and with $c(i, i) = 1$.

As noticed by Deville [4], it follows from Theorem 5.2 that many of the second-order inclusion probabilities are zero. As a result, no unbiased variance estimator may be found for the Horvitz–Thompson estimator. The search for variance estimators under reasonable model assumptions for the variable of interest y is a matter for further research.

6. Interest of ordered pivotal sampling

This is clear from Theorems 5.1 and 5.2 that ordered pivotal sampling induces a sampling design with a rather small entropy, since the second-order inclusion probabilities heavily depend on the order of the units in the population. If the maximization of entropy is a major concern, *randomized pivotal sampling*, where the list of the units in the population is randomly ordered before applying the pivotal method, should certainly be preferred. The main interest of ordered pivotal sampling lies in the gain of precision obtained from a stratification effect, if the ranking of the units in the population is well correlated to the variable of interest. In this sense, ordered pivotal sampling is similar in spirit to classical, ordered systematic sampling. However, systematic sampling can be particularly inefficient if the ordering is unappropriate, with regard to the variable of interest. Ordered pivotal sampling introduces more randomization in the sampling process, and should be more robust, in some sense, than systematic sampling. In the sequel, ordered pivotal sampling is compared to other sampling designs with respect to various criteria.

To fix ideas, we consider the case of (i) equal inclusion probabilities $\pi_k = n/N$, such that (ii) the population size N is an integer multiple of the sample size n, and we note $N = np$. In this case, the microstrata $U_i, i = 1, \dots, n$, are non overlapping with the same size $N_i = p$. We have

$$U_i = \{(i-1)p + 1, \dots, (i-1)p + p\}, \tag{6.1}$$

and ordered pivotal sampling amounts to stratified simple random sampling of size $n_i = 1$ inside each microstratum U_i. Also, it is well known that under the same assumptions (i) and (ii), systematic sampling amounts to simple random sampling of size $m = 1$ in the population $G_c = \{g_1, \dots, g_p\}$ of $M = p$ clusters, where each cluster

$$g_j = \{j, j + p, \dots, j + (n-1)p\} \tag{6.2}$$

contains $M_j = n$ units. Let y denote some variable of interest, and let

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{6.3}$$

denote the Horvitz–Thompson (HT) estimator of the total $t_y = \sum_{k \in U} y_k$.

Under conditions (i) and (ii), ordered systematic sampling and ordered pivotal sampling may be seen as particular cases of Markov chain designs (Breidt [2]). Let M be a doubly stochastic transition probability matrix, of size p. In a Markov chain design with matrix of transition M, a sample $s = \{R_1, p + R_2, \dots, (n-1)p + R_n\}$ is selected, where $R_1, \dots, R_n$ is the Markov chain associated to M, with R_1 being uniformly distributed on $\{1, \dots, p\}$. Let $I(p)$ denote the identity matrix of size p, and $J(p)$ denote the square matrix of size p with all elements equal to 1. The use of the matrix of transition

$$M_\rho = \rho \frac{J(p)}{p} + (1-\rho) I(p),$$

with $\rho \in [0, 1]$ defines the category of compromise Markov chain designs (Breidt [2]). The choice $\rho = 0$ leads to ordered systematic sampling, while the choice $\rho = 1$ leads to ordered pivotal sampling.

6.1. Entropies of sampling designs

As a measure of randomness of a sampling design $q(\cdot)$, we use the entropy $H(q)$ defined as

$$H(q) = -\sum_{s \subset U} q(s) \log q(s), \tag{6.4}$$

with $0 \log 0 = 0$ by convention. We have

$$\begin{aligned} H(srs) &= \log N! - \log n! - \log(N-n)! \\ &= \sum_{k=0}^{n-1} \log\left(\frac{N-k}{n-k}\right) \end{aligned}$$

for simple random sampling, and

$$H(sys) = \log\left(\frac{N}{n}\right)$$

for ordered systematic sampling, see for example Tillé and Haziza [14]. Some straightforward algebra leads to

$$H(ops) = n\log\left(\frac{N}{n}\right)$$

for ordered pivotal sampling. As a measure of comparison of entropy for two sampling designs $q(\cdot)$ and $r(\cdot)$, we may use the Kullback–Leibler divergence

$$D(q\|r) = \sum_{s\subset U} q(s)\log\frac{q(s)}{r(s)}$$

if the two sampling designs are such that $r(s) = 0 \Rightarrow q(s) = 0$. We obtain

$$D(sys\|srs) = \sum_{k=1}^{n-1}\log\left(\frac{N-k}{n-k}\right),$$

$$D(ops\|srs) = \sum_{k=0}^{n-1}\log\left(\frac{1-k/N}{1-k/n}\right),$$

$$D(sys\|ops) = (n-1)\log\left(\frac{N}{n}\right).$$

Both simple random sampling and ordered pivotal sampling clearly have much larger entropy than ordered systematic sampling.

6.2. Maximum design-effect for sampling designs

This is a standard fact that the variance of the HT-estimator under without-replacement simple random sampling is given by

$$V_{srs}(\hat{t}_{y\pi}) = N^2\frac{1-f}{n}S_y^2, \tag{6.5}$$

where $f = n/N$, $S_y^2 = \frac{1}{N-1}\sum_{k\in U}(y_k-\mu_y)^2$ and $\mu_y = \frac{1}{N}\sum_{k\in U} y_k$. On the other hand, the variance of the HT-estimator under ordered pivotal sampling and assumptions (i) and (ii) may be written as

$$V_{ops}(\hat{t}_{y\pi}) = N^2\frac{1-f}{n}\frac{1}{n}\sum_{i=1}^{n} S_{yi}^2, \tag{6.6}$$

where $S_{yi}^2 = \frac{1}{N_i - 1}\sum_{k \in U_i}(y_k - \mu_{yi})^2$ and $\mu_{yi} = \frac{1}{N_i}\sum_{k \in U_i} y_k$. Finally, the variance of the HT-estimator under systematic sampling is then given by

$$V_{sys}(\hat{t}_{y\pi}) = N^2 \frac{1-f}{n}\frac{1}{n} S_Y^2, \tag{6.7}$$

where

$$\begin{aligned} S_Y^2 &= \frac{1}{M-1}\sum_{j=1}^{p}\left(t_{yj} - \frac{t_y}{M}\right)^2 \\ &= \frac{n^2}{p-1}\sum_{j=1}^{p}(m_{yj} - \mu_y)^2, \end{aligned}$$

with $t_{yj} = \sum_{k \in G_j} y_k$ and $m_{yj} = t_{yj}/n$.

As a measure of risk of a strategy combining a sampling design $q(\cdot)$ and HT-estimation, we may use the maximum design-effect

$$DMAX(q) = \max_{y \in \mathcal{C}} \frac{V_q(\hat{t}_{y\pi})}{V_{srs}(\hat{t}_{y\pi})}, \tag{6.8}$$

where $\mathcal{C}$ denotes the set of non-constant variables of interest (that is, containing all variables y such that $S_y^2 \neq 0$).

Theorem 6.1. *Assume that conditions* (i) *and* (ii) *are satisfied. Then we have for ordered pivotal sampling*

$$DMAX(ops) = \frac{N-1}{N-n} \tag{6.9}$$

and for ordered systematic sampling

$$DMAX(sys) = n\frac{N-1}{N-n}. \tag{6.10}$$

Proof. For any variable y, it follows from a standard analysis of variance that

$$S_y^2 = \sum_{i=1}^{n}\frac{p-1}{N-1}S_{yi}^2 + \sum_{i=1}^{n}\frac{p}{N-1}(\mu_{yi} - \mu_y)^2,$$

so that

$$\sum_{i=1}^{n} S_{yi}^2 \leq \frac{N-1}{p-1}S_y^2$$

and the equality occurs if all the stratum means μ_{yi} are equal. A joint application of (6.5) and (6.6) leads to

$$\frac{V_{ops}(\hat{t}_{y\pi})}{V_{srs}(\hat{t}_{y\pi})} \le \frac{N-1}{n(p-1)} S_y^2 / S_y^2 = \frac{N-1}{N-n},$$

which gives (6.9). The use of an alternative analysis of variance leads to

$$S_y^2 = \sum_{j=1}^{p} \frac{n-1}{N-1} \sigma_{yj}^2 + \sum_{j=1}^{p} \frac{n}{N-1} (m_{yj} - \mu_y)^2,$$

where $\sigma_{yj}^2 = \frac{1}{n-1} \sum_{k \in G_j} (y_k - m_{yj})^2$. This leads to

$$S_Y^2 \le \frac{n^2}{p-1} \frac{N-1}{n} S_y^2,$$

and the equality occurs if the variable y is constant inside any cluster g_j. By a joint application of (6.5) and (6.7), we have

$$\frac{V_{sys}(\hat{t}_{y\pi})}{V_{srs}(\hat{t}_{y\pi})} \le \frac{n^2}{p-1} \frac{N-1}{n^2} S_y^2 / S_y^2 = n \frac{N-1}{N-n},$$

which gives (6.10). □

If the sample size n remains small to moderate, equation (6.9) implies that *DMAX* tends to 1 in case of ordered pivotal sampling, if N is sufficiently large. Even in the worst cases, ordered pivotal sampling will thus be competitive to simple random sampling. On the other hand, equation (6.10) implies that a strategy involving systematic sampling may be considerably more risky in some situations.

6.3. Equality of treatment of variables for sampling designs

As pointed out by a referee, the *DMAX* criterion considered previously is very stringent since referring to the worst possible variable for sampling designs. In case of ordered systematic sampling, this would be a cyclical variable whose period is equal to $p = N/n$; such a situation is usually unlikely to occur, except in particular situations.

An alternative criterion studied by Deville [3] and Qualité [12] considers the equality of treatment of variables. For any sampling design $q(\cdot)$ with first-order inclusion probabilities π_k and second-order inclusion probabilities $\pi_{kl}(q)$ with $k, l \in U$, let

$$\Delta(q) = [\Delta_{kl}(q)]_{k,l \in U}$$

be its design variance–covariance matrix, with $\Delta_{kl}(q) = \pi_{kl}(q) - \pi_k \pi_l$. Let $0 \le \lambda_1(q) \le \cdots \le \lambda_N(q)$ denote the eigenvalues of $\Delta(q)$. For any variable y that lies on the unit sphere of the

Euclidean norm (that is, such that $\sum_{k\in U} y_k^2 = 1$), we have

$$\lambda_1(q) \leq V_p(\hat{t}_{y\pi}) \leq \lambda_N(q).$$

Roughly speaking, the extreme eigenvalues give the extreme possible values for the variance. Qualité [12], page 50, proposed to measure the equality of treatment for variables in terms of minimization of the dispersion of the eigenvalues, denoted as

$$\delta(q) = N^{-1} \sum_{k=1}^{N} \{\lambda_k(q) - \bar{\lambda}(q)\}^2$$

with $\bar{\lambda}(q) = N^{-1} \sum_{k=1}^{N} \lambda_k(q)$. Note that $\sum_{k=1}^{N} \lambda_k(q) = Tr(\Delta(q))$, where $Tr(\cdot)$ denotes the trace. For any sampling design $q(\cdot)$ with equal probabilities $\pi_k = n/N$, this leads to

$$\begin{aligned}\bar{\lambda}(q) &= N^{-1} \sum_{k\in U} \pi_k(1-\pi_k) \\ &= \frac{p-1}{p^2},\end{aligned}$$

which will be simply denoted as $\bar{\lambda}$ in the sequel.

The ranking of the evaluated sampling designs with respect to this criterion is established in Theorem 6.2. Clearly, ordered pivotal sampling tends to treat the variables more equally than ordered systematic sampling. To demonstrate Theorem 6.2, we need the following lemma.

Lemma 6.1. *Assume that the sampling design* $q(\cdot)$ *is performed with equal probabilities* $\pi_k = n/N$, *and that conditions* (i) *and* (ii) *are satisfied. Assume that* $\Delta(q)$ *has only two eigenvalues* 0 *and* $\lambda_+(q) > 0$, *with multiplicities* $N_0(q)$ *and* $N - N_0(q)$, *respectively. Then*:

$$\delta(q) = \frac{N_0(q)}{N - N_0(q)} \bar{\lambda}^2. \tag{6.11}$$

Proof. Since $q(\cdot)$ has only one strictly positive eigenvalue $\lambda_+(q)$ with multiplicity $N - N_0(q)$, we have $\lambda_+(q) = \frac{N}{N-N_0(q)}\bar{\lambda}$. Then

$$\begin{aligned}N\delta(q) &= \sum_{k=1}^{N} \{\lambda_k(q) - \bar{\lambda}\}^2 \\ &= N_0(q)\bar{\lambda}^2 + \{N - N_0(q)\}\left\{\frac{N}{N - N_0(q)}\bar{\lambda} - \bar{\lambda}\right\}^2 \\ &= N\frac{N_0(q)}{N - N_0(q)}\bar{\lambda}^2.\end{aligned}$$

□

Theorem 6.2. *Assume that conditions* (i) *and* (ii) *are satisfied. Then we have*

$$\delta(srs) \le \delta(ops) \le \delta(sys). \tag{6.12}$$

Proof. It may be easily shown (see, e.g., Deville [3], page 120) that in case of simple random sampling, $\Delta(srs)$ has only two eigenvalues 0 and $\lambda_+(srs) = \frac{N\bar{\lambda}}{N-1}$, with $N_0(srs) = 1$.

Let $I(n)$ denotes the identity matrix of size n, and $J(n)$ denotes the square matrix of size n with all elements equal to 1. After some algebra, we obtain from the definition of ordered pivotal sampling that

$$\Delta(ops) = \begin{pmatrix} \Delta_1 & 0 & \cdots & 0 \\ 0 & \Delta_1 & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Delta_1 \end{pmatrix} = I(n) \otimes \Delta_1,$$

where $\Delta_1 = p^{-1}\{I(p) - p^{-1}J(p)\}$ and $\otimes$ denotes the Kronecker product. It follows that the N eigenvalues of $\Delta(ops)$ are given by the products of the eigenvalues of $I(n)$ and Δ_1 (see for example Theorem 1 in Magnus and Neudecker [10], page 28). The eigenvalues of Δ_1 are p^{-1} and 0 with multiplicities $p-1$ and 1, respectively. Consequently, $\Delta(ops)$ has only two eigenvalues 0 and $\lambda_+(ops) = p^{-1}$, with $N_0(ops) = n$.

Similarly, we obtain from the definition of ordered systematic sampling that

$$\Delta(sys) = \begin{pmatrix} \Delta_1 & \cdots & \cdots & \Delta_1 \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \Delta_1 & \cdots & \cdots & \Delta_1 \end{pmatrix} = J(n) \otimes \Delta_1.$$

Since the eigenvalues of $J(n)$ are 0 and n with multiplicities $n-1$ and 1, respectively, $\Delta(sys)$ has only two eigenvalues 0 and $\lambda_+(sys) = np^{-1}$, with $N_0(sys) = N - p + 1$.

Equation (6.11) implies that $\delta(q)$ increases as $N_0(q)$ increases. Clearly, $N_0(srs) \le N_0(ops)$, and from the identity $(N - p + 1) - n = (p-1)(n-1) \ge 0$, we obtain that $N_0(ops) \le N_0(sys)$ so that the result follows. □

6.4. Some numerical results on a small population

To investigate on the properties of considered sampling algorithms, we considered a small example. We first generated a finite population of size $N = 12$, containing three variables of interest, y_1, y_2 and y_3. Table 2 shows the values for the three variables of interest. The variable y_1 is highly correlated to the order of the units in the population, on the contrary to variable y_2. The variable y_3 exhibits a particularly unfavorable case for systematic sampling.

We considered equal probability sampling of size $n = 2$ (respectively, $n = 4$) by means of six sampling designs: simple random sampling without replacement (SRS), ordered systematic sampling (SYS), compromise Markov chain design with $\rho = 0.25$ (CMC25), $\rho = 0.50$ (CMC50),

Table 2. Values of three variables of interest in the generated population

Unit	1	2	3	4	5	6	7	8	9	10	11	12
y_1	10	10	10	15	45	45	50	50	60	60	60	65
y_2	15	45	10	60	60	50	45	65	10	50	10	60
y_3	10	45	60	15	50	65	10	50	60	10	45	60

$\rho = 0.75$ (CMC75), and ordered pivotal sampling (OPS). As a measure of variability of the HT-estimator $\hat{t}_{y\pi}$ for a sampling design $q(\cdot)$, we considered the design-effect (DEFF) given by

$$DEFF = \frac{V_q(\hat{t}_{y\pi})}{V_{srs}(\hat{t}_{y\pi})}, \tag{6.13}$$

where the variances are computed by means of formulas (6.5)–(6.7) for SRS, OPS and SYS, and from formula (1) in Breidt [2], page 66, for compromise Markov chain designs. Table 3 shows DEFF for the five strategies. As could be expected, the CMC25, CMC50 and CMC75 give compromise results between SYS and OPS. Also, it is clear from Table 3 that both OPS and SYS lead to a subsequent reduction of variance for variable y_1, with $DEFF$ ranging from 0.17 to 0.50 and OPS performing significantly better. The OPS strategy is essentially similar to SRS for the variable y_2 which is poorly correlated to the order of the units in the population, while SYS may be much worse ($DEFF = 1.39$ for $n = 2$) or much better ($DEFF = 0.36$ for $n = 4$). Finally, we obtain for the variable y_3 a considerable loss for SYS, while the loss is more limited for OPS with $DEFF = 1.10$ for $n = 2$ and $DEFF = 1.36$ for $n = 4$.

Acknowledgements

The author wishes to thank two referees for useful comments and suggestions that helped improving the quality of the paper significantly.

Table 3. Design-effect for three variables of interest and five strategies in the generated population

	Sample size $n = 2$			Sample size $n = 4$		
	y_1	y_2	y_3	y_1	y_2	y_3
SYS	0.50	1.39	2.18	0.27	0.36	5.44
CMC25	0.46	1.31	1.91	0.24	0.61	3.94
CMC50	0.43	1.24	1.64	0.21	0.76	2.81
CMC75	0.39	1.17	1.37	0.19	0.85	1.97
OPS	0.35	1.10	1.10	0.17	0.95	1.36

References

[1] Bondesson, L. and Grafström, A. (2011). An extension of Sampford's method for unequal probability sampling. *Scand. J. Stat.* **38** 377–392. MR2829606

[2] Breidt, F.J. (1995). Markov chain designs for one-per-stratum sampling. *Survey Methodology* **21** 63–70.

[3] Deville, J.-C. (1987). Les sondages complexes. In *Les Sondages* 89–136. Paris: Economica.

[4] Deville, J.-C. (1998). Une nouvelle méthode de tirage à probabilités inégales. Technical Report 9804, Ensai, France.

[5] Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85** 89–101. MR1627234

[6] Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91** 893–912. MR2126040

[7] Deville, J.-C. and Tillé, Y. (2005). Variance approximation under balanced sampling. *J. Statist. Plann. Inference* **128** 569–591. MR2102777

[8] Godinot, A. (2005). Pour comprendre le Recensement de la population. *Séries Insee Méthodes, Hors Série*. Paris: INSEE.

[9] Jay Breidt, F. and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *J. Statist. Plann. Inference* **141** 479–487. MR2719512

[10] Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Chichester: Wiley. MR0940471

[11] Nedyalkova, D., Qualité, L. and Tillé, Y. (2009). General framework for the rotation of units in repeated survey sampling. *Stat. Neerl.* **63** 269–293. MR2751893

[12] Qualité, L. (2010). Unequal probability sampling and repeated surveys. Ph.D. dissertation, Univ. Neuchâtel, Switzerland.

[13] Tillé, Y. (2006). *Sampling Algorithms. Springer Series in Statistics*. New York: Springer. MR2225036

[14] Tillé, Y. and Haziza, D. (2010). An interesting property of the entropy of some sampling designs. *Survey Methodology* **36** 229–231.

Received November 2010 and revised April 2011