INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES
Série des Documents de Travail du CREST
(Centre de Recherche en Economie et Statistique)

n° 2010-50

On a Characterization of Ordered Pivotal Sampling

G. CHAUVET[1]

1. CREST-ENSAI. chauvet@ensai.fr

On a characterization of ordered pivotal sampling

G. CHAUVET *

October 27, 2010

RESUME

Lorsqu'une information auxiliaire est disponible à l'étape du plan de sondage, il est possible de sélectionner un échantillon par tirage équilibré à l'aide de la méthode du Cube (Deville et Tillé, 2004). Nous nous intéressons ici à un cas particulier de cet algorithme, appelé la méthode du pivot (Deville et Tillé, 1998). Nous montrons que cet algorithme est équivalent à la méthode systématique de Deville, au sens où les deux algorithmes mettent en oeuvre le même plan de sondage. Cette caractérisation permet le calcul des probabilités d'inclusion doubles. Nous montrons également que la méthode du pivot permet d'utiliser un tri préalable des unités afin d'obtenir une réduction de variance, tout en limitant la perte d'efficacité si la variable de tri n'est pas explicative de la variable d'intérêt.

ABSTRACT

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.

*Crest(Ensai), chauvet@ensai.fr

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 (2004) 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 (2005) 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 (2011). The second assumption is related to the entropy of the sampling design: the variance approximations proposed by Deville and Tillé (2005) 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 the 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, Census surveys

are conducted annually; the detailed methodology is described in Godinot (2005). 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 (2005) to hold are clearly not respected.

We are interested in a particular case of the cube method, called *pivotal sampling* (Deville and Tillé, 1998), 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 (1998), and known in the literature as *Deville's systematic sampling* (Tillé, 2006). 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 (1998) may be readily applied to pivotal sampling. This provides an answer to a problem raised by Bondesson and Grafström (2010, p. 7). Deville's systematic sampling has found uses in the context of longitudinal surveys, see Nedyalkova et al. (2009).

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 $k = 1, \ldots, N$. 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 $\pi = (\pi_1, \ldots, \pi_N)'$. 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 non-negative integer i. The cross-border units are denoted as $k_i,\ i = 1, \ldots, n-1$, and we note $a_i = i - V_{k_i - 1}$ and $b_i = V_{k_i} - i$. The microstratum $U_i,\ i = 1, \ldots, n$, is defined as

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

with $k_0 = 0$ and $k_n = N + 1$. The microstrata are generally overlapping, since one cross-border unit may belong to two adjacent microstrata. In the particular case when $V_{k_i} = i$, which implies that $b_i = 0$, we consider equivalently that the cross-border unit k_i only belongs to the microstratum U_i, or that it belongs also

to the microstratum U_{i+1} as a phantom unit. To fix ideas, useful quantities for population U are presented in Fig. 1.

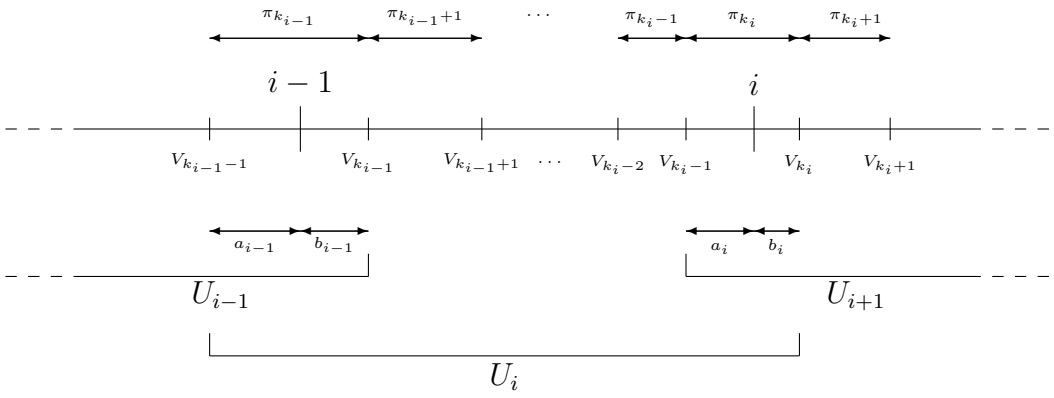

Figure 1: Inclusion probabilities and cross-border units in microstratum U_i, for population U

The N sampling units are grouped to obtain a population $U_c = \{u_1, \ldots, u_{2n-1}\}$ of clusters as follows: for $i = 1, \ldots, n$, we define

$$u_{2i-1} = \{k_{i-1} + 1, \ldots, k_i - 1\} \tag{2}$$

to which is associated the probability $\phi_{2i-1} = V_{k_i-1} - V_{k_{i-1}}$. In the particular case when $V_{k_i-1} = V_{k_{i-1}}$, we consider u_{2i-1} as a phantom cluster with probability $\phi_{2i-1} = 0$. For $i = 1, \ldots, n-1$, we define

$$u_{2i} = \{k_i\} \tag{3}$$

to which is associated the probability $\phi_{2i} = V_{k_i} - V_{k_i-1} = \pi_{k_i}$. We note $\psi = (\phi_1, \ldots, \phi_{2n-1})'$. To fix ideas, useful quantities for population U_c are presented in Fig. 2.

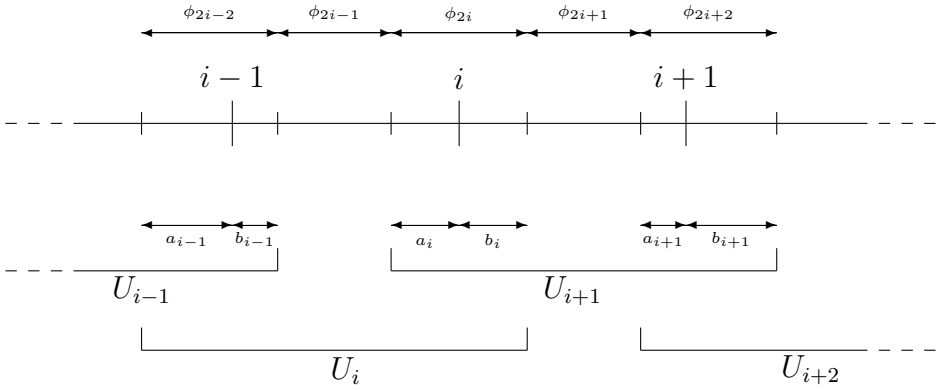

Figure 2: Inclusion probabilities and cross-border units in microstrata U_i and U_{i+1} for population U_c

3 Ordered pivotal sampling

A general algorithm for pivotal sampling is described in Deville and Tillé (1998). 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 more 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. Let $J_0 = 1$. 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 loser is definitely eliminated while the survivor gets the sum of their probabilities and then faces the next unit. The fights go on until the accumulated probability exceeds 1, which occurs when the cross-border unit k_i is involved. One of the two remaining units, denoted as W_i, wins and is then definitely selected in the sample while the other one, denoted as J_i, jumps to the following microstratum. Lemma 1 states that Algorithm 1 may alternatively be seen as a two-stage procedure. The proof follows from definition, and is thus omitted.

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

1. We initialize with $\pi(0) = \pi$.
2. At step t:
   (a) Let $k < l$ be the two units in U with the smaller indexes such that $0 < \pi_k(t-1), \pi_l(t-1) < 1$.
   (b) If $m \in U \setminus \{k, l\}$, then $\pi_m(t) = \pi_m(t-1)$.
   (c) If $\pi_k(t-1) + \pi_l(t-1) < 1$, let

$$[\pi_k(t), \pi_l(t)] = \begin{cases} [\pi_k(t-1) + \pi_l(t-1), 0] & \text{with prob. } \lambda_1(t) \\ [0, \pi_k(t-1) + \pi_l(t-1)] & \text{with prob. } 1 - \lambda_1(t) \end{cases}$$

   where

$$\lambda_1(t) = \frac{\pi_k(t-1)}{\pi_k(t-1) + \pi_l(t-1)},$$

   and otherwise, let

$$[\pi_k(t), \pi_l(t)] = \begin{cases} [1, \pi_k(t-1) + \pi_l(t-1) - 1] & \text{with prob. } \lambda_1(t) \\ [\pi_k(t-1) + \pi_l(t-1) - 1, 1] & \text{with prob. } 1 - \lambda_1(t) \end{cases}$$

   where

$$\lambda_1(t) = \frac{1 - \pi_l(t-1)}{2 - \pi_k(t-1) - \pi_l(t-1)}.$$

3. The procedure ends at step T, when $\pi(T)$ has only integer (0-1) components.

Lemma 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.*

We assume that a sample S_{op} is selected in U_c by means of ordered pivotal sampling with parameter ψ, and we let $X_1 < \ldots < X_n$ denote the units selected in the sample, ranked in ascending order. Lemma 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 gives the probabilities for the different outcomes in the case of a non cross-border unit u_{2i-1}.

Lemma 2 *In case of ordered pivotal sampling with parameter ψ, we have*

$$\{X_i = u_{2i-2}\} \Rightarrow \{J_{i-1} \in \{X_1, \ldots, X_i\}\}, \tag{4}$$

$$\{X_i = u_{2i-1}\} \Rightarrow \{W_i = u_{2i-1}\} \cup \{J_i = u_{2i-1}\}, \tag{5}$$

$$\{X_i = u_{2i}\} \Rightarrow \{J_i \notin \{X_1, \ldots, X_i\}\}. \tag{6}$$

Proof of lemma. *Assume that $X_i = u_{2i-2}$. This implies that i units exactly are selected in the $i-1$ first microstrata $U_1, \ldots, U_{i-1}$. On the other hand, if $J_{i-1} \notin \{X_1, \ldots, X_i\}$ the unit J_{i-1} is not selected in the sample so that at most $i-1$ units are selected in $U_1, \ldots, U_{i-1}$. This proves (4), and by a similar argument we obtain (6). It is easily seen that (5) 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 *In case of ordered pivotal sampling with parameter ψ, we have*

$$pr\,(W_i = u_{2i-1}) = \frac{(1-a_i-b_{i-1})(1-a_i-b_i)}{(1-a_i)(1-b_i)}, \tag{7}$$

$$pr\,(J_i = u_{2i-1}) = \frac{a_i(1-a_i-b_{i-1})}{(1-a_i)(1-b_i)}, \tag{8}$$

$$pr\left(X_i = u_{2i-1}\right) \;=\; 1 - a_i - b_{i-1}. \tag{9}$$

Proof of lemma. *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 :

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

which gives (7). Similarly, we obtain

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

which gives (8). We now consider equation (9). Since

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

and

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

it suffices to show that

$$\{u_{2i-1} \in S_{op}\} \Rightarrow \{X_i = u_{2i-1}\}. \tag{10}$$

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 $J_{i-1} \notin \{X_1, \ldots, X_i\}$. In other words, $\{u_{2i-1} \in S_{op}\}$ implies that exactly $i - 1$ units smaller than u_{2i-1} were selected, which proves (10).

Finally, let $U_{c,i} = \{u_{2i-2}, \ldots, u_{2n-1}\}$, $\psi_i = (b_{i-1}, \phi_{2i-1}, \ldots, \phi_j, \ldots, \phi_{2n-1})'$, and $S_{op,i}$ be a random sample selected in $U_{c,i}$ by means of ordered pivotal sampling with parameter ψ_i. Lemma 4 establishes some relations for conditional inclusion probabilities in $S_{op,i}$ of the first units in $U_{c,i}$.

Lemma 4

$$\begin{aligned} & pr\left(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i} \mid u_{2i-2} \in S_{op,i}\right) \\ = \; & \frac{b_i}{1-a_i}, \end{aligned} \tag{11}$$

$$\begin{aligned} & pr\left(u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} \mid u_{2i-2} \in S_{op,i}\right) \\ = \; & \frac{(1-a_i-b_i)(1-b_i-a_{i+1})}{(1-a_i)(1-b_i)}, \end{aligned} \tag{12}$$

$$\begin{aligned} & pr\left(u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i} \mid u_{2i-2} \in S_{op,i}\right) \\ = \; & \frac{(1-a_i-b_i)a_{i+1}}{(1-a_i)(1-b_i)}. \end{aligned} \tag{13}$$

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

We first consider equation (11). Since b_{i-1} is the first-order inclusion probability of unit u_{2i-2} in sample $S_{op,i}$, we have

$$pr\left(u_{2i-2} \in S_{op,i}\right) \;=\; b_{i-1}. \tag{14}$$

On the other hand, the event

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

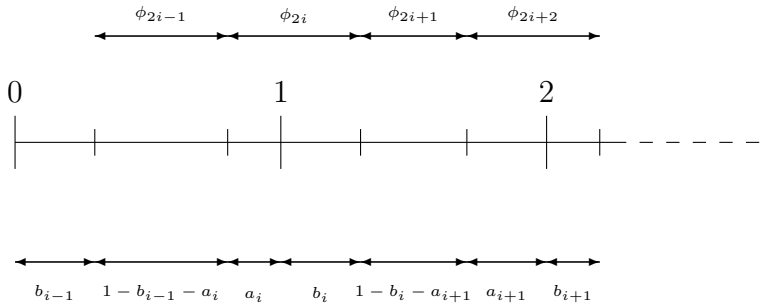

Figure 3: Inclusion probabilities and cross-border units in the two first microstrata of population $U_{c,i}$

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:

$$\begin{aligned} & pr\left(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\right) \\ = \ & \frac{b_{i-1}}{1-a_i} \times 1 \times b_i, \end{aligned} \tag{15}$$

and equation (11) follows from (14) and (15). We now consider equation (12). 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:

$$\begin{aligned} & pr\left(u_{2i+1} \in S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\right) \\ = & \frac{b_{i-1}}{1-a_i} \times \frac{1-a_i-b_i}{1-b_i} \times \frac{1-b_i-a_{i+1}}{1-a_{i+1}} \times (1-a_{i+1}), \\ = & \frac{b_{i-1}(1-a_i-b_i)(1-b_i-a_{i+1})}{(1-a_i)(1-b_i)}, \end{aligned} \tag{16}$$

which, together with (14), leads to (12). Finally, we consider equation (13). 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:

$$\begin{aligned} & pr\left(u_{2i+2} \in S_{op,i}, u_{2i+1} \notin S_{op,i}, u_{2i} \notin S_{op,i}, u_{2i-1} \notin S_{op,i}, u_{2i-2} \in S_{op,i}\right) \\ = & \frac{b_{i-1}}{1-a_i} \times \frac{1-a_i-b_i}{1-b_i} \times 1 \times \frac{a_{i+1}}{1-b_{i+1}} \times (1-b_{i+1}), \\ = & \frac{b_{i-1}(1-a_i-b_i)a_{i+1}}{(1-a_i)(1-b_i)}, \end{aligned} \tag{17}$$

which gives (13).

4 Deville's systematic sampling

The sampling algorithm known in the literature as Deville's systematic sampling (Deville, 1998; Tillé, 2006) 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 (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 5 follows from the definition of Algorithm 2.

Lemma 5 *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

$$pr\left(Y_{i+1} = u_j \mid Y_1, \ldots, Y_i\right) = pr\left(Y_{i+1} = u_j \mid Y_i\right). \tag{18}$$

The different cases for the transition probabilities in (18) easily follow from the definition of Algorithm 2, and are given below:

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 $V_{k-1} \leq w_1 < V_k$.

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 $a_{i-1}b_{i-1}\left\{(1-a_{i-1})(1-b_{i-1})\right\}^{-1}$,
   - according to a distributed Uniform$(0,1)$ random variable with probability $1-a_{i-1}b_{i-1}\left\{(1-a_{i-1})(1-b_{i-1})\right\}^{-1}$.
2. The unit k is selected if $V_{k-1} \leq w_i + (i-1) < V_k$.

$$
\begin{aligned}
& pr\left(Y_{i+1}=u_j \mid Y_1,\ldots,Y_{i-1},Y_i=u_{2i-2}\right) \\
= & \begin{cases} \frac{b_i}{1-a_i} & (j=2i), \\ \frac{(1-b_i-a_{i+1})(1-a_i-b_i)}{(1-a_i)(1-b_i)} & (j=2i+1), \\ \frac{a_{i+1}(1-a_i-b_i)}{(1-a_i)(1-b_i)} & (j=2i+2). \end{cases}
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
& pr\left(Y_{i+1}=u_j \mid Y_1,\ldots,Y_{i-1},Y_i=u_{2i-1}\right) \\
= & \begin{cases} \frac{b_i}{1-a_i} & (j=2i), \\ \frac{(1-b_i-a_{i+1})(1-a_i-b_i)}{(1-a_i)(1-b_i)} & (j=2i+1), \\ \frac{a_{i+1}(1-a_i-b_i)}{(1-a_i)(1-b_i)} & (j=2i+2). \end{cases}
\end{aligned} \tag{20}
$$

$$
\begin{aligned}
& pr\left(Y_{i+1} = u_j \mid Y_1, \ldots, Y_{i-1}, Y_i = u_{2i}\right) \\
= & \begin{cases} \frac{(1-b_i-a_{i+1})}{(1-b_i)} & (j = 2i+1), \\ \frac{a_{i+1}}{(1-b_i)} & (j = 2i+2). \end{cases}
\end{aligned} \tag{21}
$$

5 Second-order inclusion probabilities

We can now formulate our main result.

Theorem 1 *Ordered pivotal sampling and Deville's systematic sampling with the same parameter π induce the same sampling design.*

Proof of theorem. *From Lemmas 1 and 5, 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 (19)-(21) hold in case of ordered pivotal sampling. Recall that we note*

$$
\begin{aligned}
U_{c,i} &= \{u_{2i-2}, \ldots, u_{2n-1}\}, \\
\psi_i &= \left(b_{i-1}, \phi_{2i-1}, \ldots, \phi_j, \ldots, \phi_{2n-1}\right)',
\end{aligned}
$$

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 (19). From (4), we obtain:

$$
\begin{aligned}
& pr\left(X_{i+1} = u_{2i} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-2}\right) \\
= \; & pr\left(X_{i+1} = u_{2i} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-2}, J_{i-1} \in \{X_1, \ldots, X_i\}\right),
\end{aligned}
$$

which is equivalent to $pr\left(u_{2i} \in S_{op,i}, u_{2i-1} \notin S_{op,i} \mid u_{2i-2} \in S_{op,i}\right)$, so that the result follows from equation (11).

Similarly, we obtain

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

where the last line follows from (12), and

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

where the last line follows from (13). This proves equation (19). The proof for equation (21) is similar, and is thus omitted.

We now turn to equation (20). We introduce some further notation. Let

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

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

$$\begin{aligned} & pr\left(X_{i+1} = u_{2i} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-1}, W_i = u_{2i-1}\right) \\ = \;& pr\left(X_{i+1} = u_{2i} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-1}, J_i = u_{2i}\right) \end{aligned}$$

$$\equiv \quad pr\left(u_{2i} \in S_{op,i+1}\right) = b_i, \tag{22}$$

where the second line in (22) comes from

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

Also,

$$pr\left(X_{i+1} = u_{2i} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-1}, J_i = u_{2i-1}\right) \quad = \quad 1, \tag{23}$$

since

$$\begin{aligned} \{X_i = u_{2i-1}, J_i = u_{2i-1}\} \quad &\Rightarrow \quad \{X_i = u_{2i-1}, W_i = u_{2i}\} \\ &\Rightarrow \quad \{X_{i+1} = u_{2i}\}. \end{aligned}$$

Further,

$$\begin{aligned} & pr\left(W_i = u_{2i-1} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-1}\right) \\ = \quad & pr\left(W_i = u_{2i-1} \mid X_i = u_{2i-1}\right) \\ = \quad & pr\left(X_i = u_{2i-1} \mid W_i = u_{2i-1}\right) \frac{pr\left(W_i = u_{2i-1}\right)}{pr\left(X_i = u_{2i-1}\right)} \\ = \quad & 1 \times \frac{(1 - a_i - b_{i-1})(1 - a_i - b_i)\left\{(1 - a_i)(1 - b_i)\right\}^{-1}}{1 - a_i - b_{i-1}} \\ = \quad & \frac{1 - a_i - b_i}{(1 - a_i)(1 - b_i)}, \end{aligned} \tag{24}$$

the fourth line in (24) being a consequence of Lemma 3. The same reasoning leads to

$$pr\left(J_i = u_{2i-1} \mid X_1, \ldots, X_{i-1}, X_i = u_{2i-1}\right)$$

$$
\begin{aligned}
&= pr\left(J_i = u_{2i-1} \mid X_i = u_{2i-1}\right) \\
&= pr\left(X_i = u_{2i-1} \mid J_i = u_{2i-1}\right) \frac{pr\left(J_i = u_{2i-1}\right)}{pr\left(X_i = u_{2i-1}\right)} \qquad (25) \\
&= b_i \times \frac{a_i(1-a_i-b_{i-1})\left\{(1-a_i)(1-b_i)\right\}^{-1}}{1-a_i-b_{i-1}} \\
&= \frac{a_i b_i}{(1-a_i)(1-b_i)}.
\end{aligned}
$$

From equations (22)-(25), we obtain that

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

Similar computations lead to

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

and

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

which proves (20).

Theorem 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 (1998), and is reminded below.

Theorem 2 (Deville, 1998) *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

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

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

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

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

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

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

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

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

As noticed by Deville (1998), it follows from Theorem 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 1 and 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.

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 $N = n\ p$. In this case, the microstrata $U_i,\ i = 1, \ldots, n$, are non overlapping with the same size $N_i = p$. We have

$$U_i = \{(i-1)p + 1, \ldots, (i-1)p + p\}, \tag{26}$$

and ordered pivotal sampling amounts to stratified simple random sampling of size $n_i = 1$ inside each microstratum U_i. Let y denote some variable of interest, and let

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{27}$$

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

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

$$V_{srs}\left(\hat{t}_{y\pi}\right) = N^2 \, \frac{1-f}{n} \, S_y^2, \tag{28}$$

where $f = n/N$, $S_y^2 = \frac{1}{N-1} \sum_{k \in U} \left(y_k - \mu_y\right)^2$ 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

$$V_{ops}\left(\hat{t}_{y\pi}\right) = N^2 \, \frac{1-f}{n} \, \frac{1}{n} \sum_{i=1}^{n} S_{yi}^2, \tag{29}$$

where $S_{yi}^2 = \frac{1}{N_i - 1} \sum_{k \in U_i} \left(y_k - \mu_{yi}\right)^2$ and $\mu_{yi} = \frac{1}{N_i} \sum_{k \in U_i} y_k$. Finally, 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 $G_c = \{g_1, \ldots, g_p\}$ of $M = p$ clusters, where each cluster

$$g_j = \{j, j+p, \ldots, j+(n-1)p\} \tag{30}$$

contains $M_j = n$ units. The variance of the HT-estimator under systematic sampling is then given by

$$V_{sys}\left(\hat{t}_{y\pi}\right) = N^2 \, \frac{1-f}{n} \, \frac{1}{n} S_Y^2, \tag{31}$$

where

$$S_Y^2 \;=\; \frac{1}{M-1} \sum_{j=1}^{p} \left(t_{yj} - \frac{t_y}{M} \right)^2$$

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

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 $p(\cdot)$ and HT-estimation, we use the maximum design-effect

$$DMAX(p) = \max_{y \in \mathcal{C}} \frac{V_p\left(\hat{t}_{y\pi}\right)}{V_{srs}\left(\hat{t}_{y\pi}\right)}, \tag{32}$$

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

$$DMAX(ops) = \frac{N-1}{N-n} \tag{33}$$

and for ordered systematic sampling

$$DMAX(sys) = n\,\frac{N-1}{N-n}. \tag{34}$$

Proof of theorem. *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 (28) and (29) leads to

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

which gives (33). The use of an alternative analysis of variance leads to

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

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

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

and the equality occurs if the variable y is constant inside any cluster g_j. By a joint application of (28) and (31), we have

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

which gives (33).

If the sample size n remains small to moderate, equation (33) 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 (34) implies that a strategy involving systematic sampling may be considerably more risky in some situations.

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 1 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.

Table 1: 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

We considered equal probability sampling of size $n = 2$ (respectively, $n = 4$) by means of (i) simple random sampling without replacement (SRS), (ii) ordered pivotal sampling (OPS), and (iii) ordered systematic sampling (SYS). As a measure of variability of the HT-estimator $\hat{t}_{y\pi}$ for a sampling design $p(\cdot)$, we considered the design-effect (DEFF) given by

$$DEFF = \frac{V_p\left(\hat{t}_{y\pi}\right)}{V_{srs}\left(\hat{t}_{y\pi}\right)}, \tag{35}$$

where the variances are computed by means of formulas (28), (29) and (31). Table 2 shows DEFF for strategies OPS and SYS. It is clear from Table 2 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 far 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 far much limited for OPS with $DEFF = 1.10$ for $n = 2$ and $DEFF = 1.36$ for $n = 4$.

Table 2: Design-effect for three variables of interest and two 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
OPS	0.35	1.10	1.10	0.17	0.95	1.36
SYS	0.50	1.39	2.18	0.27	0.36	5.44

References

Bondesson, L., and Grafström, A. (2010). *An extension of Sampford's method for unequal probability sampling*. To appear in Scandinavian Journal of Statistics.

Breidt, F.J., and Chauvet, G. (2011). *Improved variance estimation for balanced samples drawn via the Cube method*. Journal of Statistical Planning and Inference, 141, pages 479-487.

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

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

Deville, J-C., and Tillé, Y. (2004). *Efficient balanced sampling: the cube method*. Biometrika, 91, pages 893-912.

Deville, J-C., and Tillé, Y. (2005). *Variance approximation under balanced sampling*. Journal of Statistical Planning and Inference, 128, pages 569-591.

Godinot, A. (2005). *Pour comprendre le Recensement de la population*. Séries Insee Méthodes, Hors Série.

Nedyalkova, D., and Qualité, L., and Tillé, Yves (2009). *General Framework for the Rotation of Units in Repeated Survey Sampling*. Statistica Neerlandica, 63, pages 269-293.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Series in Statistics, New-York.