

UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT THROUGH A SPLITTING METHOD

J.-C. Deville and Y. Tillé¹

ABSTRACT

A very general class of sampling methods without replacement and with unequal probabilities is proposed. It consists in splitting the inclusion probability vector into several new inclusion probability vectors. One of these vectors is chosen randomly; thus, the initial problem is reduced to another sampling problem with unequal probabilities. This operation is repeated and, at each step, the sampling problem is reduced to a simpler problem. Any sampling design can be implemented by means of the splitting technique. The simplicity of this technique allows the easy generation of new sampling procedures with unequal probabilities. First, a sampling procedure with only N samples with non-null probabilities is presented. A method of splitting into simple random samplings is also proposed. The Midzuno method, which can be described as a splitting procedure, is generalised in such a way that it can be applied to any inclusion probability vector. The elimination procedure can also be presented as a particular case of the splitting method. It is also shown that the Chao procedure can be presented as a strict elimination procedure. Next a very simple method based on the modification of two inclusion probabilities is defined. A sufficient condition in order that a splitting method satisfy the Sen-Yates-Grundy conditions is given. Finally we discuss the Gabler sufficient condition, which ensures that a sampling design without replacement always yields estimators with a smaller variance than in the case with replacement. It is shown that the elimination procedure satisfies the Gabler condition.

RÉSUMÉ

Les auteurs proposent une catégorie très générale de méthodes d'échantillonnage à tirage exhaustif et à probabilités inégales. Ces méthodes consistent à diviser le vecteur des probabilités d'inclusion en plusieurs nouveaux vecteurs. On choisit ensuite l'un d'eux au hasard, et ainsi, le problème initial est réduit à un autre problème d'échantillonnage à probabilités inégales. On reprend l'opération, et à chaque étape, le problème d'échantillonnage se simplifie. Cette technique de division autorise l'application de n'importe quel plan d'échantillonnage. Sa simplicité permet d'engendrer aisément de nouvelles méthodes d'échantillonnage à probabilités inégales. Les auteurs présentent d'abord une méthode d'échantillonnage comprenant seulement N échantillons à probabilité non nulle. Ils proposent aussi une méthode de division débouchant sur un échantillonnage simple aléatoire. La méthode de Midzuno, qui est en sorte une technique de division, est généralisée de manière à s'appliquer à n'importe quel vecteur des probabilités d'inclusion. La méthode par élimination peut aussi être considérée comme un cas particulier de la technique de division. Il en va autant de la méthode de Chao, dans laquelle on peut voir l'utilisation stricte de la méthode par élimination. Les auteurs définissent ensuite une méthode très simple reposant sur la modification de deux probabilités d'inclusion. Ils fixent une condition suffisante pour que la technique de division satisfasse les conditions de Sen-Yates-Grundy. Enfin, ils parlent de la condition suffisante de Gabler, en vertu de laquelle un plan d'échantillonnage à tirage exhaustif donne toujours des estimateurs à variance plus faible que les plans à tirage non exhaustif. La méthode par élimination satisfait la condition de Gabler.

1. INTRODUCTION

Consider a finite population $U = \{1, \dots, k, \dots, N\}$ whose units can be identified by an order number. For each unit k , it is supposed that the value y_k of the characteristic y can be measured. The objective is to

estimate the total

$$t_y = \sum_{k \in U} y_k,$$

by means of a random sample without replacement and of fixed sample size n . Suppose also that the values $x_k > 0$ of an auxiliary characteristic x are known for all

¹ Jean-Claude Deville, Unité des Méthodes Statistiques, Institut National de Statistique et des Études Économiques, 18, Blvd. Adolphe Pinard, 75675 Paris Cédex, France, E-mail: Jean-Claude.Deville@DG75E.insee.atlas.fr; 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

units of U and that the x_k are approximately proportional to the y_k . It is then interesting to select the units with unequal probabilities in order to get an estimator of t_y with a small variance.

To implement such a sampling design, the inclusion probabilities π_k of each unit are computed as follows. First define the function

$$h(z) = \sum_{k \in U} \min \left(z \frac{x_k}{t_x}, 1 \right), \quad (1)$$

where $t_x = \sum_{k \in U} x_k$. The inclusion probabilities are given by

$$\pi_k = \min \left(1, h^{-1}(n) \frac{x_k}{t_x} \right). \quad (2)$$

Since the selection of the units selected with an inclusion probability equalling one is trivial, it is supposed that the problem consists in selecting n units without replacement in a population of size N with fixed inclusion probabilities $\pi_k, k \in U$, where $0 < \pi_k < 1, k \in U$, and

$$\sum_{k \in U} \pi_k = n.$$

A sample s is a non-empty subset of U and a sampling design $p(s)$ of fixed sample size n is a probability distribution on all the possible samples of size n such that

$$p(s) \geq 0 \text{ and } \sum_{s \in S} p(s) = 1,$$

where

$$S = \{s \in U \mid \#s = n\}.$$

Moreover, a sampling design which respects fixed inclusion probabilities must satisfy the relations

$$\sum_{s \ni k} p(s) = \pi_k, k \in U.$$

The random sample is denoted S and is such that $Pr(S=s)=p(s)$. The π -estimator (see Horvitz and Thompson, 1952) allows the estimation without bias of the total

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

When the design is of fixed sample size n , the variance of this estimator can be written

$$V(\hat{t}_{y\pi}) = \sum_{k \in U} \sum_{\ell \in U} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 (\pi_k \pi_\ell - \pi_{k\ell}),$$

and can be estimated by

$$\hat{V}(\hat{t}_{y\pi}) = \sum_{k \in S} \sum_{\ell \in S} \left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell} \right)^2 \frac{\pi_k \pi_\ell - \pi_{k\ell}}{\pi_{k\ell}},$$

where $\pi_{k\ell}$ is the probability to select jointly units k and ℓ and is called the joint inclusion probability.

Remember that the sufficient condition in order that $\hat{V}(\hat{t}_{y\pi})$ be unbiased is that $\pi_{k\ell} > 0, k \neq \ell \in U$. Yates and Grundy (1953) and Sen (1953) however showed that if $\pi_{k\ell} \leq \pi_k \pi_\ell, k \neq \ell \in U$, then this variance estimator always takes a positive value. This sufficient condition is called the Sen-Yates-Grundy condition. An important property of a sampling procedure is that the variance of the π -estimator should be always better than the variance obtained by sampling with replacement. This property was only proved for two sampling designs: Sampford's procedure (Gabler, 1981, 1984) and Chao's procedure (Sengupta, 1989). Besides these two properties, a sampling procedure must present some practical interest. It should be possible to apply a "good" method without enumerating the $N! \{n!(N-n)!\}^{-1}$ samples of size n . Moreover, the joint inclusion probabilities should also be easily computable.

Since the publication of the synthesis book of Brewer and Hanif (1983), which presents fifty sampling procedures with unequal probabilities, several new methods have been proposed. One can cite Chao's sequential procedure whose properties were studied by Bethlehem and Schuerhoff (1984) and Sengupta (1989). The method of emptying boxes (see Hedayat, Lin and Stufken, 1989, and Hedayat and Sinha, 1991) provides a very general frame applicable to any rational inclusion probabilities. Finally, the elimination procedure (see Tillé, 1996) is a simple algorithm related to Chao's (1973) and Sinha's (1973) methods. The elimination method is also related to the Midzuno method (see Brewer and Hanif, 1983, p. 25), the generalisation of which will be discussed.

The proposed method is directly inspired from the technique of emptying boxes. Nevertheless it is generalised to any non-rational inclusion probabilities. After the presentation of the splitting method, it is shown that any sampling design can be defined by means of this technique (section 2). Moreover, the splitting technique allows the easy conception of new sampling procedures with unequal probabilities. First, two new simple sampling procedures with unequal probabilities are proposed in sections 3 and 4. The first one provides the minimal support design and the second breaks up the problem into simple random samplings. Next, it is shown that the splitting

technique allows the easy definition of the Midzuno (section 5), elimination (section 6) and Chao (section 7) procedures. The Midzuno procedure is however generalised to be applicable to any inclusion probability vector. The presentation of these procedures by means of the splitting technique shows that they belong to the same family of methods. The pivotal method, which is interesting for its extreme simplicity, is also defined in section 8. In this method, only two inclusion probabilities are modified at each step. Its implementation is almost as simple as systematic sampling.

The splitting procedure is also interesting because it allows the discussion of general properties for a large set of sampling methods. In section 8, a general sufficient condition is given in order that the Sen-Yates-Grundy condition be satisfied. This condition is directly applicable to the generalised Midzuno method, the elimination method, the Chao procedure and the pivotal method. Finally, the Gabler (1984) sufficient condition is discussed in section 10. This condition ensures that a sampling design without replacement is always better than sampling with replacement (see Hansen and Hurwitz, 1943). Several counter-examples are given to show that the Gabler condition is not satisfied by most of the methods discussed. Nevertheless, it is shown that the elimination procedure satisfies this condition and thus always gives a better estimator than sampling with replacement.

2. THE SPLITTING METHOD

The basic technique is extremely simple: it consists in splitting the π_k into two parts $\pi_k^{(1)}$ and $\pi_k^{(2)}$, $k \in U$. The $\pi_k^{(1)}$ and $\pi_k^{(2)}$, $k \in U$, can take any values that satisfy the following relations:

$$\pi_k = \lambda \pi_k^{(1)} + (1 - \lambda) \pi_k^{(2)}, \quad k \in U, \quad (3)$$

$$0 \leq \pi_k^{(1)} \leq 1 \quad \text{and} \quad 0 \leq \pi_k^{(2)} \leq 1, \quad k \in U, \quad (4)$$

where λ can be chosen freely but must be such that $0 < \lambda < 1$, and

$$\sum_{k \in U} \pi_k^{(1)} = \sum_{k \in U} \pi_k^{(2)} = n. \quad (5)$$

The method consists in drawing n units with unequal probabilities

$$\begin{cases} \pi_k^{(1)}, & k \in U, \quad \text{with a probability } \lambda \\ \pi_k^{(2)}, & k \in U, \quad \text{with a probability } 1 - \lambda. \end{cases}$$

The problem is thus reduced to another sampling problem with unequal probabilities. If the splitting is such that one or several of the $\pi_k^{(1)}$ and of the $\pi_k^{(2)}$ are equal to 0 or 1, the sampling problem will be simpler at the next step. If this operation is repeated, the problem can be reduced until the sample is obtained. The splitting can also be organised in order that one of the vectors of $\pi_k^{(1)}$ or of $\pi_k^{(2)}$ has equal probabilities. In this case, a simple random sampling without replacement can be applied directly.

This technique of splitting can be generalised to a splitting technique into M vectors of inclusion probabilities that will be denoted by an element of $A \subset U$, where $\#A = M$. In order to apply this method, we first build the $\pi_k^{(j)}$, $j \in A$, and the λ_j , $j \in A$, in such a way that

$$\sum_{j \in A} \lambda_j = 1,$$

$$\sum_{j \in A} \lambda_j \pi_k^{(j)} = \pi_k, \quad k \in U,$$

$$0 \leq \pi_k^{(j)} \leq 1, \quad k \in U, j \in A,$$

and

$$\sum_{k \in U} \pi_k^{(j)} = n, \quad j \in A.$$

The method simply consists in selecting one of the vectors of $\pi_k^{(j)}$ with unequal probabilities λ_j , $j \in A$. Again, the $\pi_k^{(j)}$ are built in such a way that the sampling problem will be simpler at the next step. Note that, when a unit k is such that π_k equals one or zero, it means that this unit is definitively drawn or eliminated from the sample. Indeed, if $\pi_k = 0$ then $\pi_k^{(j)} = 0$ for all $j \in A$ and if $\pi_k = 1$ then $\pi_k^{(j)} = 1$ for all $j \in A$.

Any sampling method $p(s)$ can be implemented by means of the splitting technique. A sampling design can always be presented as a tree each of whose leaves corresponds to a sample s with non-null probabilities $p(s)$. By gathering branches of this tree, it is thus possible to build an infinity of representation by means of the splitting technique of a given sampling design.

For instance, from any sampling design, a sampling method into two parts can be built by using a unit i from U and by taking $\lambda_1 = \pi_i$, $\lambda_2 = 1 - \pi_i$. Next, define

$$\pi_k^{(1)} = \begin{cases} \frac{\pi_{ki}}{\pi_i} & \text{if } k \neq i \\ 1 & \text{if not} \end{cases}$$

and

$$\pi_k^{(2)} = \begin{cases} \frac{\pi_k - \pi_{ki}}{1 - \pi_i} & \text{if } k \neq i \\ 0 & \text{if not.} \end{cases}$$

At the next step, third order inclusion probabilities must be used, which makes the procedure difficult to apply in practice. Another way to define a splitting method is to use the implementation in sampling schemes such as defined by Hedayat et Sinha (1991, pp. 7-11). Nevertheless, these two techniques are complex and need the knowledge of all the $p(s)$. Rather than implementing a sampling design from the known $p(s)$, it seems more interesting to present several splitting techniques that are easy to implement, and next to discuss their properties.

3. MINIMAL SUPPORT DESIGN

It is always possible to define a sampling design that respects any fixed inclusion probabilities by using only N samples with non-null probabilities. This result, which follows from theorem 1 of Wynn (1977), was not presented with a procedure of construction of this design. This is however the most obvious application of the splitting technique into two parts and is related to the simplest "game" proposed by Hedayat, Lin and Stufken (1989, theorem 2.1.) for the sampling procedure through a method of emptying boxes.

In order to present this method, first note $\pi_{(1)}, \dots, \pi_{(k)}, \dots, \pi_{(N)}$, the ordered inclusion probabilities. Next, define

$$\lambda = \min\{1 - \pi_{(N-n)}, \pi_{(N-n+1)}\},$$

$$\pi_{(k)}^{(1)} = \begin{cases} 0 & \text{if } k \leq N-n \\ 1 & \text{if } k > N-n, \end{cases}$$

and

$$\pi_{(k)}^{(2)} = \begin{cases} \frac{\pi_k}{1 - \lambda} & \text{if } k \leq N-n \\ \frac{\pi_k - \lambda}{1 - \lambda} & \text{if } k > N-n. \end{cases}$$

With a probability λ , the sample of the n units that have the largest inclusion probabilities is selected, and, with a probability $1 - \lambda$, another sample of size n must

be selected with the inclusion probabilities $\pi_k^{(2)}$. However, one of the $\pi_k^{(2)}$ equals zero or one. Indeed, if $1 - \pi_{(N-n)} < \pi_{(N-n+1)}$ then $\lambda = 1 - \pi_{(N-n)}$ and $\pi_{(N-n)}^{(2)} = 1$. On the other hand, if $\pi_{(N-n+1)} < 1 - \pi_{(N-n)}$ then $\lambda = \pi_{(N-n+1)}$ and $\pi_{(N-n)}^{(2)} = 0$. The problem is thus reduced to the selection of a sample in a population of size $N-1$.

Example 1 Suppose that $N = 6, n = 3, \pi_1 = 0.07, \pi_2 = 0.17, \pi_3 = 0.41, \pi_4 = 0.61, \pi_5 = 0.83, \pi_6 = 0.91$. In this case, the breakdown is completed in 4 steps. The vector of inclusion probabilities is split into two parts given in columns 2 and 3 of table 1. With a probability $\lambda = 0.59$, units 4, 5, 6 are selected and with a probability $1 - \lambda = 0.41$, another sampling with unequal probabilities given by (0.171 0.415 1 0.049 0.585 0.780) is applied. At step 2, the splitting is applied again of this last vector and in 4 steps at the most, the sample is selected.

Table 1

π_k	Step 1 $\lambda = 0.59$	Sept 2 $\lambda = 0.585$	Step 3 $\lambda = 0.471$	Step 4 $\lambda = 0.778$
0.07	0 0.171	0 0.412	0 0.778	1 0
0.17	0 0.415	0 1	1 1	1 1
0.41	0 1	1 1	1 1	1 1
0.61	1 0.049	0 0.118	0 0.222	0 1
0.83	1 0.585	1 0	0 0	0 0
0.91	1 0.780	1 0.471	1 0	0 0

The sampling design is thus given by $p(\{4, 5, 6\}) = 0.59, p(\{3, 5, 6\}) = (1 - 0.59) \times 0.585 = 0.24, p(\{2, 3, 6\}) = (1 - 0.59 - 0.24) \times 0.471 = 0.08, p(\{1, 2, 3\}) = (1 - 0.59 - 0.24 - 0.08) \times 0.778 = 0.07, p(\{2, 3, 4\}) = (1 - 0.59 - 0.24 - 0.08 - 0.7) = 0.02$.

4. SPLITTING INTO SIMPLE RANDOM SAMPLINGS

This method also consists in splitting the inclusion probabilities into two parts. In this case, the sampling problem is reduced to N simple random samplings without replacement. Again, the same notation will be used for the ordered inclusion probabilities $\pi_{(1)}, \dots, \pi_{(k)}, \dots, \pi_{(N)}$. Next, define

$$\lambda = \min \left\{ \pi_{(1)} \frac{N}{n}, \frac{N}{N-n} (1 - \pi_{(N)}) \right\},$$

Table 2

π_k	Step 1 $\lambda = 0.14$		Step 2 $\lambda = 0.058$		Step 3 $\lambda = 0.173$		Step 4 $\lambda = 0.045$		Step 5 $\lambda = 0.688$	
0.07	0.5	0	0	0	0	0	0	0	0	0
0.17	0.5	0.116	0.600	0.086	0.5	0	0	0	0	0
0.41	0.5	0.395	0.600	0.383	0.5	0.358	0.667	0.344	0.5	0
0.61	0.5	0.628	0.600	0.630	0.5	0.657	0.667	0.656	0.5	1
0.83	0.5	0.884	0.600	0.901	0.5	0.985	0.667	1	1	1
0.91	0.5	0.977	0.600	1	1	1	1	1	1	1

and compute

$$\pi_{(k)}^{(1)} = \frac{n}{N}, k \in U,$$

and

$$\pi_{(k)}^{(2)} = \frac{\pi_k - \lambda \frac{n}{N}}{1 - \lambda}, k \in U.$$

If $\lambda = \pi_{(1)} N/n$, then $\pi_{(1)}^{(2)} = 0$. On the other hand, if $\lambda = (1 - \pi_{(N)}) N/(N - n)$ then $\pi_{(N)}^{(2)} = 1$. At the next step, the problem is thus reduced to a selection of a sample of size $n - 1$ or n from a population of size $N - 1$. In N steps at the most, the problem can be solved.

Example 2 If the same inclusion probabilities are used as in example 1, the result of the splitting method into simple random samplings is given in table 2.

The problem consists in selecting one of the 6 simple random sampling designs defined by the columns of table 3. The simple random sampling design are selected with a probability given in the upper margin of table 3.

Table 3

k	Probabilities					
	0.14	0.050	0.14	0.03	0.44	0.200
1	0.5	0	0	0	0	0
2	0.5	0.6	0.5	0	0	0
3	0.5	0.6	0.5	0.667	0.5	0
4	0.5	0.6	0.5	0.667	0.5	1
5	0.5	0.6	0.5	0.667	1	1
6	0.5	0.6	1	1	1	1

This breakdown into simple random sampling designs also yields a quite general result.

Theorem 1 For any vector $(\pi_1, \dots, \pi_k, \dots, \pi_N)'$, such that $0 < \pi_k < 1, k \in U$, and

$$\sum_{k \in U} \pi_k = n,$$

there always exists at least one sampling design such that

$$p(s) > 0, s \in S$$

Proof

The splitting method into simple random samplings implies that, with a probability λ , a single random sampling is applied on U as follows:

$$p(s) \geq \left(\frac{N}{n}\right)^{-1} \min \left\{ \pi_{(1)} \frac{N}{n}, \frac{N}{N-n} (1 - \pi_{(N)}) \right\} > 0.$$

5. GENERALIZED MIDZUNO METHOD

The Midzuno method (see Horvitz and Thompson, 1952) is another way to split the inclusion probabilities into simple random sampling designs without replacement. This method can be described as a splitting procedure into N parts. First, define

$$\lambda_j = \pi_j \frac{N-1}{N-n} - \frac{n-1}{N-n}, j \in U.$$

Since λ_j must be in $[0, 1]$, the method is applicable only if

$$\pi_k \geq \frac{n-1}{N-1}, k \in U. \tag{6}$$

This condition is very restrictive. For this reason, the Midzuno method has no great practical interest. This implementation is however quite simple and can easily be described as a splitting method into N parts.

One of the vectors of $\pi^{(j)}$, $j \in U$, will be selected with the probabilities λ_j . The $\pi^{(j)}$ are the inclusion probabilities by means of which the selection will be applied at the next step where

$$\pi_k^{(j)} = \begin{cases} 1 & \text{if } k=j \\ \frac{n-1}{N-1} & \text{if } k \neq j. \end{cases}$$

Since, at the second step, the problem is reduced to sampling with equal probabilities (except one unit that is selected automatically), a simple random sampling is applied. The joint inclusion probabilities are not difficult to derive (see for instance, Brewer and Hanif, 1983, p.25) and are given by

$$\pi_{kt} = \frac{n-1}{N-2} \left(\pi_k + \pi_t - \frac{n}{N-1} \right). \quad (7)$$

By (6), we directly get

$$\pi_{kt} \geq \frac{(n-1)(n-2)}{(N-1)(N-2)} > 0.$$

The joint inclusion probabilities are thus strictly positive. It is shown in section 9 that the Midzuno method satisfies de Sen-Yates-Grundy condition.

The Midzuno method can be generalized in order that it can be applied to any inclusion probabilities even if condition (6) is not satisfied. The quantities λ_j are computed by means of the following algorithm. First, define $F := \emptyset$. Next, repeat until getting a stable configuration of the $\lambda_j, j \in U$, the two following allocations:

$$\lambda_j := \begin{cases} 1 - \frac{(1-\pi_j)(N-1-\#F)}{N-n-\sum_{i \in F}(1-\pi_i)} & \text{if } j \notin F \\ 0 & \text{if } j \in F, \end{cases}$$

$$F := \{j \in U \mid \lambda_j \leq 0\}.$$

Finally, define

$$\pi_k^{(j)} = \begin{cases} 1 & \text{if } k=j \\ \frac{\pi_k - \lambda_k}{1 - \lambda_k} & \text{if } k \neq j. \end{cases}$$

The validity of the method can be easily verified. Moreover, when only one iteration is needed, the algorithm provides the λ_j of Midzuno's method. The fundamental difference with the classical Midzuno method is that the problem is not necessarily reduced to a simple random sampling at the second step. The algorithm is thus repeated until a simple random sampling is obtained.

Example 3 Again, if we take the same data as in examples 1 and 2, at the first step, $\lambda_i = 0$, $i = 1, \dots, 4$, $\lambda_5 = 0.346$ et $\lambda_6 = 0.654$. The problem is thus reduced to a splitting into two parts (see table 4).

At the second step, the method provides a splitting of each of these two parts into 3 parts. Finally, we get the breakdown given in table 5.

At step 3, the problem consists in selecting only one unit. This example however, shows that the generalized Midzuno method does not ensure strictly positive joint inclusion probabilities; indeed $\pi_{12} = 0$.

Table 4

$\lambda_5 = 0.346$	$\lambda_6 = 0.654$
0.07	0.07
0.17	0.17
0.41	0.41
0.61	0.61
1	0.74
0.74	1

Table 5

0.017	0.128	0.201	0.032	0.243	0.380
0.07	0.07	0.07	0.07	0.07	0.07
0.17	0.17	0.17	0.17	0.17	0.17
1	0.38	0.38	1	0.38	0.38
0.38	1	0.38	0.38	1	0.38
1	1	1	0.38	0.38	1
0.38	0.38	1	1	1	1

6. ELIMINATION PROCEDURE

In order to implement the elimination procedure (see Tillé, 1996), the inclusion probabilities $\pi(k|i)$, $k \in U$, for the sample sizes $i = n, \dots, N$ are computed as follows:

$$\pi(k|i) = \min \left\{ 1, h^{-1}(i) \frac{x_k}{I_x} \right\}, k \in U, i = n, \dots, N, \quad (8)$$

where $h(\cdot)$ is defined in (1). Note the two particular cases: $\pi(k|n) = \pi_k$ and $\pi(k|N) = 1, k \in U$.

The steps of the algorithm are numbered in a decreasing order from $N - 1$ to n . At each step of the procedure, a unit is eliminated from U with a probability

$$r_{ki} = 1 - \frac{\pi(k|i)}{\pi(k|i+1)}, k \in U, i = n, \dots, N-1. \quad (9)$$

After $N - n$ steps, a sample of n units is selected. The inclusion probability of a unit is thus the probability that this unit has not been eliminated during the $N - n$ steps. We thus obtain the relation:

$$\pi_k = \prod_{i=n}^{N-1} (1 - r_{ik}). \quad (10)$$

The joint inclusion probability of unit k and ℓ is the probability that neither unit k nor ℓ has been eliminated during the $N - n$ steps and thus

$$\pi_{k\ell} = \prod_{i=n}^{N-1} (1 - r_{ik} - r_{i\ell}). \quad (11)$$

The elimination method can also be presented as a splitting technique. The computation of the λ_j is also given by an algorithm. First, define $F := \emptyset$. Next, repeat until getting a stable configuration of the $\lambda_j, j \in U$, the two following allocations:

$$\lambda_j := \begin{cases} 1 - \pi_j \frac{N-1-\#F}{n - \sum_{i \in F} (1 - \pi_i)} & \text{if } j \notin F \\ 0 & \text{if } j \in F. \end{cases}$$

$$F := \{j \in U | \lambda_j \leq 0\}.$$

Finally define

$$\pi_k^{(j)} = \begin{cases} 0 & \text{if } k=j \\ \frac{\pi_k}{1 - \lambda_k} & \text{if } k \neq j. \end{cases}$$

The complementary design $p^c(s)$ of a sampling design $p(s)$ is defined as the sampling design, such that the sample $s^c = U \setminus s$ is selected with a probability $p^c(s)$. If the design $p(s)$ has the inclusion probabilities π_k , then $p^c(U \setminus s) = p(s)$, $\pi_k^c = 1 - \pi_k$, $k \in U$ and $\pi_{k\ell}^c = 1 - \pi_k - \pi_\ell + \pi_{k\ell}$. It is easy to see that the generalised Midzuno method is the complementary method of the elimination method. The joint inclusion probabilities of the Midzuno method can thus be derived from those of the elimination method.

Example 4 Again, if we use the data of example 1, we obtain at the first step the splitting presented in table 6.

Table 6

$\lambda_1 = 0.708333$	$\lambda_2 = 0.291667$
0	0.24
0.24	0
0.41	0.41
0.61	0.61
0.83	0.83
0.91	0.91

At the first step, we already see that π_{12} equals zero. The splitting obtained at the second step is presented in table 7.

At this step, 3 units must be selected amongst the 4 remaining units, which is obvious.

7. CHAO'S METHOD

This method, such as presented by Chao (1982), is a sequential updating procedure that seems not to be related with the elimination method. This can however be defined as an elimination procedure. Indeed, at each of the $N - n$ steps of the algorithm, a unit is definitively eliminated from the sample. Chao's procedure differs from the elimination method because in the Chao procedure, a unit is eliminated from the $n + 1$ first units of the sample. The Chao method can thus be presented as a splitting technique into $M = n + 1$ parts. We have thus $A = \{1, \dots, n+1\}$ and, again, the probabilities $\lambda_j, j \in A$, are given by an algorithm. First define $F := \emptyset$. Next, repeat the two following allocations until a stable configuration for the $\lambda_j, j \in A$, is obtained:

Table 7

0.438492	0.247354	0.0224867	0.180556	0.101852	0.00925926
0	0	0	0	0.63	0.63
0	0.63	0.63	0	0	0
0.63	0	0	0.63	0	0.63
0.63	0.63	0.63	0.63	0.63	0
0.83	0.83	0.83	0.83	0.83	0.83
0.91	0.91	0.91	0.91	0.91	0.91

$$\lambda_j := \begin{cases} 1 - \pi_j \frac{n - \#F}{\sum_{i \in A \setminus F} \pi_i} & \text{if } j \notin F \\ 0 & \text{if } j \in F, \end{cases}$$

$$F := \{j \in A \mid \lambda_j \leq 0\}.$$

Next define

$$\pi_k^{(j)} = \begin{cases} 0 & \text{if } k = j \\ \frac{\pi_k}{1 - \lambda_k} & \text{if } k \neq j, k \in A \\ \pi_k & \text{if } k \notin A. \end{cases}$$

At the next step, repeat the same algorithm on the first $n + 1$ remaining units. Remember that, with Chao's method, the result obtained depends on the order of the units. With the data of example 1 sorted in increasing order, we obtain, in this case, the same result as in table 6.

Again, this presentation can be generalised. Indeed, it is possible to apply the same method by eliminating one unit amongst any $A \subset U, \#A > n$. The algorithm for computing the λ_j is identical as for the Chao method. The elimination procedure is the case where $A = U$. At the next step, a unit amongst another subset of U can be eliminated. This subset must not necessarily have the same size as A .

8. PIVOTAL METHOD

The pivotal method is based on a splitting into two parts of the vector of inclusion probabilities. Only two inclusion probabilities are modified. The method consists in selecting two units that will be denoted i and j .

If $\pi_i + \pi_j > 1$, then

$$\lambda = \frac{1 - \pi_j}{2 - \pi_i - \pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ 1 & k = i \\ \pi_i + \pi_j - 1 & k = j, \end{cases}$$

and

$$\pi_k^{(2)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ \pi_i + \pi_j - 1 & k = i \\ 1 & k = j. \end{cases}$$

On the other hand, if $\pi_i + \pi_j < 1$, then

$$\lambda = \frac{\pi_i}{\pi_i + \pi_j},$$

$$\pi_k^{(1)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ \pi_i + \pi_j & k = i \\ 0 & k = j, \end{cases}$$

and

$$\pi_k^{(2)} = \begin{cases} \pi_k & k \in U \setminus \{i, j\} \\ 0 & k = i \\ \pi_i + \pi_j & k = j. \end{cases}$$

In the first case, a one is allocated to only one inclusion probability. In the second case, a zero is allocated to only one inclusion probability. The problem is thus reduced to a population of size $N - 1$. In N steps at most, a solution is thus obtained. This

method is interesting for its extreme simplicity. The pivotal procedure can be implemented by means of a strictly sequential procedure, *i.e.*, it can be applied in only one reading of a data file. The sequential procedure is presented in algorithm 1 by means of the Mathematica® syntax.

Algorithm 1

```

NN=Length[pi];
i=1; a=pi[[1]];
Do[b=pi[[j]];
  If[a+b<=1,
    If[Random[]<=b/(a+b), i=j];a=a+b,
    If[Random[]<=(1-b)/(2-a-b),
      Print [i];i=j, Print[i] ]; a=a+b-1
    ],
  {j, 2, NN } ];
Print[i];

```

In algorithm 1, “pi” is the vector of inclusion probabilities. The method is thus extremely simple but unfortunately, it has an inconvenience: most of the joint inclusion probabilities equal zero. In order to overcome this problem, the procedure can be randomised (as for systematic sampling) by sorting randomly the data file. We shall see in section 9, that this method has however an advantage on systematic sampling because, for the pivotal method, the Sen-Yates-Grundy condition is satisfied.

9. SEN-YATES-GRUNDY CONDITION

Several proposed methods satisfy the Sen-Yates-Grundy condition ($\pi_{k\ell} \leq \pi_k \pi_\ell, k \neq \ell \in U$). A sufficient condition in order that the Sen-Yates-Grundy condition be satisfied is given by the following theorem.

Theorem 2 *If a sampling method without replacement is implemented by the splitting technique in M parts indexed by a subset A of U and if, at each step, the inclusion probabilities can be written as follows:*

$$\pi_k^{(j)} = \begin{cases} \gamma_k & \text{if } k \in A \setminus \{j\} \\ C & \text{if } k = j, \end{cases}$$

for all $j \in A$ where C equals 0 or 1, then this method satisfies the Sen-Yates-Grundy condition.

Proof

The proof is given by induction. At the last step, the inclusion probabilities equal 0 or 1, and, in this case, the Sen-Yates-Grundy condition is satisfied. The Sen-Yates-Grundy condition is supposed to be satisfied at the next step, *i.e.*,

$$\pi_{k\ell}^{(j)} \leq \pi_k^{(j)} \pi_\ell^{(j)}, k \neq \ell \in U, j \in A. \quad (12)$$

where $\pi_{k\ell}^{(j)}$ denotes the joint inclusion probabilities obtained by applying the splitting technique to the $\pi_k^{(j)}, j \in U$. By (12), we get

$$\pi_{k\ell} = \sum_{j \in A} \lambda_j \pi_{k\ell}^{(j)} \leq \sum_{j \in A} \lambda_j \pi_k^{(j)} \pi_\ell^{(j)}, k \neq \ell \in U. \quad (13)$$

We thus want to show that

$$\pi_{k\ell} \leq \pi_k \pi_\ell, k \neq \ell \in U, \quad (14)$$

and since

$$\pi_k = \sum_{j \in A} \lambda_j \pi_k^{(j)}, k \in U, \quad (15)$$

the expression (14) can also be written as

$$\pi_{k\ell} \leq \left(\sum_{j \in A} \lambda_j \pi_k^{(j)} \right) \left(\sum_{j \in A} \lambda_j \pi_\ell^{(j)} \right), k \neq \ell \in U. \quad (16)$$

By considering (13), a sufficient condition in order to satisfy (16) is that

$$\left(\sum_{j \in A} \lambda_j \pi_k^{(j)} \right) \left(\sum_{j \in A} \lambda_j \pi_\ell^{(j)} \right) - \sum_{j \in A} \lambda_j \pi_k^{(j)} \pi_\ell^{(j)} \geq 0, \quad (17)$$

$$k \neq \ell \in U.$$

Now,

$$\sum_{j \in A} \lambda_j \pi_k^{(j)} = \gamma_k \sum_{j \in A} \lambda_j - \lambda_k \gamma_k + \lambda_k C \quad (18)$$

$$= \gamma_k + \lambda_k (C - \gamma_k). \quad (19)$$

and

$$\begin{aligned} & \sum_{j \in A} \lambda_j \pi_k^{(j)} \pi_\ell^{(j)} \\ &= \gamma_k \gamma_\ell \sum_{j \in A} \lambda_j - \gamma_k \gamma_\ell \lambda_k - \gamma_k \gamma_\ell \lambda_\ell + \lambda_k C \gamma_\ell + \lambda_\ell C \gamma_k \\ &= \gamma_k \gamma_\ell + \lambda_k \gamma_\ell (C - \gamma_k) + \lambda_\ell \gamma_k (C - \gamma_\ell). \end{aligned} \quad (20)$$

We thus get

$$\left(\sum_{j \in A} \lambda_j \pi_k^{(j)} \right) \left(\sum_{j \in A} \lambda_j \pi_\ell^{(j)} \right) - \sum_{j \in A} \lambda_j \pi_k^{(j)} \pi_\ell^{(j)} = \lambda_k \lambda_\ell (C - \gamma_k)(C - \gamma_\ell), k \neq \ell \in U. \quad (21)$$

The condition (17) is thus always satisfied when C equals 0 or 1.

Corollary The generalised Midzuno method, the elimination method, the Chao procedure, and the pivotal method (randomised or not) satisfy the Sen-Yates-Grundy condition.

10. COMPARISON WITH SAMPLING WITH REPLACEMENT

The use of random sampling without replacement is only justified if the variance of the total estimator is smaller than when the random sampling with replacement with unequal probabilities is used. The random sampling with replacement can be described as follows: n units are selected from U independently with unequal probabilities $P_k = \pi_k/n, k \in U$. The Hansen-Hurwitz estimator (1943) of t_y is defined by

$$\hat{t}_{yHH} = \sum_{k \in U} a_k \frac{Y_k}{P_k},$$

where a_k denotes the number of times that unit k is selected in the sample. The variance of this estimator is given by

$$\text{Var}(\hat{t}_{yHH}) = \frac{1}{n} \sum_{k \in U} P_k \left(\frac{Y_k}{P_k} - t_y \right)^2.$$

Sengupta (1989) has shown that Chao's method always provides a better result than sampling with replacement.

It is also quite easy to show that a sufficient condition in order that $\text{Var}(\hat{t}_{y\pi}) \leq \text{Var}(\hat{t}_{yHH})$ for any values $\gamma_k, k \in U$, is that

$$\pi_{k\ell} \geq \frac{n-1}{n} \pi_k \pi_\ell, k \neq \ell \in U.$$

The interest in this condition is quite restricted. A much more interesting result has been given by Gabler (1984) who gives the following sufficient condition:

$$G = \sum_{k \in U} \min_{\ell \neq k} \frac{\pi_{k\ell}}{\pi_k} \geq n-1.$$

Gabler however shows that Sampford's procedure (1967) satisfies this condition.

The minimal support design no longer satisfies the Gabler condition. It is easy to build a counter-example. If $N = 4, n = 2, \pi_1 = \pi_2 = 1/3, \pi_3 = \pi_4 = 2/3$, we get $p(\{1, 2\}) = 1/3, p(\{3, 4\}) = 2/3$ and thus $G = 0$. The splitting method into simple random samplings does not satisfy the Gabler condition any more. Here is a counter-example: if $N = 4$ and $n = 2$ and that $\pi_1 = 1/4, \pi_2 = \pi_3 = 1/2, \pi_4 = 3/4$, then $G = 7/9 < n-1 = 1$. The Midzuno method does not always satisfy the Gabler condition even if $\pi_k \geq (n-1)/(N-1), k \in U$. For this case, the counter-example is more difficult to build: here is an example where $N = 20$ and $n = 3$:

$$\pi_k = \frac{3}{28}, k = 1, \dots, 5,$$

$$\pi_6 = \pi_7 = \frac{9}{70},$$

$$\pi_k = \frac{3}{20}, k = 8, \dots, 12,$$

$$\pi_k = \frac{6}{35}, k = 13, \dots, 16,$$

$$\pi_k = \frac{27}{140}, k = 17, \dots, 20.$$

In this case, by using (7), we get $G = 15152/7695 < n-1 = 2$.

We shall show that Gabler's condition is verified for the elimination method by using the following lemma.

Lemma 1 If $\pi_{k\ell}/\pi_k$ can be written as an increasing (resp. decreasing) function of the π_ℓ then the Gabler condition is satisfied.

Proof Suppose that $\pi_{k\ell}/\pi_k$ is an increasing (resp. decreasing) function of the π_ℓ and note m the label of the smallest (resp. largest) $\pi_k, k \in U$, then

$$\sum_{k \in U} \min_{\ell \neq k} \frac{\pi_{k\ell}}{\pi_k} = \sum_{k \in U} \frac{\pi_{km}}{\pi_m} + \min_{\ell \neq m} \frac{\pi_{m\ell}}{\pi_m} = n-1 + \min_{\ell \neq m} \frac{\pi_{m\ell}}{\pi_m}.$$

Theorem 3 The elimination method satisfies the Gabler condition.

Proof By the computation procedure of the $\pi(k|i)$ given by (8), it is easy to see that $\pi(k|i)/\pi(k|i+1)$ is increasing in π_k . By (9) we directly see that r_{ki} is decreasing in π_k . Next, by (10) and (11), we get

$$\frac{\pi_{k\ell}}{\pi_\ell} = \prod_{i=n}^{N-1} \frac{1-r_{ik}-r_{i\ell}}{1-r_{i\ell}} = \prod_{i=n}^{N-1} \left\{ 1 - \frac{r_{ik}}{1-r_{i\ell}} \right\}.$$

Since $r_{\ell i}$ is increasing in π_ℓ , $1-r_{ik}(1-r_{i\ell})^{-1}$ is increasing in π_ℓ and always positive. By considering that a product of positive increasing functions is always positive, the theorem is proved by the lemma 2.

ACKNOWLEDGEMENT

The authors are very grateful to Pierre Lavallée for very constructive comments that have improved this paper considerably.

REFERENCES

- Bethlehem, J.G., and Schuerhoff, M.H. (1984). "Second-order inclusion probabilities in sequential sampling without replacement with unequal probabilities", *Biometrika*, 71, 642-644.
- Brewer, K.R.W., and Hanif, M. (1983). *Sampling with Unequal Probabilities*, New York: Springer-Verlag.
- Chao, M.T. (1982). "A general purpose unequal probability sampling plan", *Biometrika*, 69, 653-656.
- Gabler, S. (1981). "A comparison of Sampford's sampling procedure versus unequal probability sampling with replacement", *Biometrika*, 68, 725-727.
- Gabler, S. (1984). "On unequal probability sampling: sufficient conditions for the superiority of sampling without replacement", *Biometrika*, 84, 171-175.
- Hanif, M., and Brewer, K.R.W. (1980). "Sampling with unequal probabilities without replacement: a review", *International Statistical Review*, 48, 217-335.
- Hansen, M.H., and Hurwitz, W.N. (1943). "On the theory of sampling from finite populations", *Annals of Mathematical Statistics*, 14, 333-362.
- Hartley, H.O., and Rao, J.N.K. (1962). "Sampling with unequal probabilities without replacement", *Annals of Mathematical Statistics*, 33, 350-374.
- Hedayat, A.S., Lin, B.-Y., and Stufken, J. (1989). "The construction of PPS sampling designs through a method of emptying boxes", *Annals of Statistics*, 17, 1886-1905.
- Hedayat, A.S., and Sinha, K.S. (1991). *Design and Inference in Finite Population Sampling*, New York: Wiley.
- Horvitz, D.G., and Thompson, D.J. (1952). "A generalisation of sampling without replacement from a finite universe", *Journal of the American Statistical Association*, 47, 663-685.
- Sampford (1967). "On sampling without replacement with unequal probabilities of selection", *Biometrika*, 54, 499-513.
- Sen, A.R. (1953). "On the estimate of variance in sampling with varying probabilities", *Journal of the Indian Society of Agricultural Statistics*, 76, 192-196.
- Sengupta, S. (1989). "On Chao's unequal probability sampling plan", *Biometrika*, 76, 192-196.
- Sinha, B.K. (1973). "On sampling schemes to realize pre-assigned sets of inclusion probabilities of first two orders", *Bulletin of the Calcutta Statistical Association*, 22, 89-110.
- Tillé, Y. (1996). "An elimination procedure for unequal probability sampling without replacement", *Biometrika*, 83, 238-241.
- Wynn, H.P. (1977). "Convex sets on finite population plans", *Annals of Statistics*, 5, 414-418.
- Yates, F., and Grundy, P.M. (1953). "Selection without replacement from within strata with probability proportional to size", *Journal of the Royal Statistical Society*, 15, 235-261.