ANNALES D'ÉCONOMIE ET DE STATISTIQUE. – N° 44 — 1996

Some remarks on unequal probability sampling designs without replacement

Yves TILLÉ*

ABSTRACT. – A set of demands is presented which should be satisfied by a good unequal-probability sampling method without replacement. First it is shown that some of these demands are contradictory. In particular, it is shown that a sequential algorithm that ensures strictly positive joint inclusion probabilities does not exist; not does there exist a sequential procedure that yields a result which is not dependent on the order of units in the data file. Next, a way is discussed to build approximations of the joint-inclusion probabilities for sampling design implemented by means of a sequential algorithm and preceded by a random sort of the data file. An original approximation is proposed which resorts to a method of adjustment to marginal totals. Finally, several approximations are compared to systematic sampling, and to Sunter's method.

Quelques remarques sur les plans de sondage sans remise à probabilités inégales

RÉSUMÉ. – Un ensemble de propriétés auxquelles devrait satisfaire un bon algorithme de tirage sans remise et à probabilités inégales est énoncé. On montre d'abord que certaines de ces propriétés sont contradictoires. Plus particulièrement, on montre qu'il n'existe pas d'algorithme de tirage séquentiel garantissant des probabilités d'inclusion d'ordre deux strictement positives. On montre également qu'il n'existe pas d'algorithme de tirage séquentiel donnant un résultat qui ne dépend pas de l'ordre du fichier. On discute ensuite de la manière de construire une approximation des probabilités d'inclusion d'ordre deux pour des plans de sondages effectués avec un algorithme séquentiel précédé d'un tri préalable du fichier de données. Une approximation originale est proposée au moyen d'une méthode de calage sur marges. Enfin, on compare diverses approximations avec le tirage systématique et la méthode de Sunter.

* Y. TILLÉ: École Nationale de la Statistique et de l'Analyse de l'Information, rue Blaise Pascal, Campus de Ker Lann, 35170 Bruz France, E-mail: tille@ensai.fr

The author is grateful to two anonymous reviewers for constructive comments which allowed to improve this paper considerably.

1 Introduction

1.1. The problem

The abundance of works about sampling designs without replacement with fixed sample size is quite disconcerting. No less than fifty sampling procedures are presented in the famous paper of HANIF and BREWER [1980]. Several other recent papers deal with this matter as well. Among them, one can cite the SUNTER method [1977 and 1986], the CHAO updating procedure [1982] and the DEVILLE method [1992].

The problem consists in drawing n units by means of a random sampling without replacement from a finite population U. This population is made up of N units which are supposed to be pointed out by an order number in such a way that it can be written: $U = \{1, ..., k, ..., N\}$. Moreover, each unit $k \in U$, must be selected with an inclusion probability fixed *a priori* π_k such that

$$(1) \qquad \sum_{k \in U} \pi_k = n \qquad \text{where} \quad 0 < \pi_k < 1, \quad k \in U.$$

More formally, the problem is generally presented as follows: a sampling design without replacement is a probability distribution $p(.)$ on all the non-empty subsets $s \subset U$ such that

$$(2) \qquad \sum_{s \subset U} p(s) = 1 \qquad \text{and} \qquad p(s) \geq 0.$$

A sampling design of fixed sample size n is such that $p(s) = 0$, if card$(s) \neq n$. An unequal probability sampling design must respect fixed first-order inclusion probability, i.e.

$$(3) \qquad \sum_{s \ni k} p(s) = \pi_k, \qquad \text{for all} \quad k \in U.$$

The use of unequal inclusion probabilities sampling can often be interesting in survey sampling. If $\pi_k > 0,\ k \in U$, the total of a variable y:

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

can be estimated without bias by the π-estimator (or HORVITZ-THOMPSON estimator) given by

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

where S is the random sample drawn by means of the sampling design $p(.)$ in such a way that $Pr(S = s) = p(s)$. If the sample size is fixed, then the variance of this estimator can be written

$$\mathrm{Var}\left(\hat{t}_{y\pi}\right) = \frac{1}{2}\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2(\pi_k\pi_\ell - \pi_{k\ell}) \tag{4}$$

where $\pi_{k\ell}$ is the joint inclusion probabilities of units k and ℓ. This variance can be estimated by

$$\widehat{\mathrm{Var}}\left(\hat{t}_{y\pi}\right) = \frac{1}{2}\sum_{k\in S}\sum_{\substack{\ell\in S\\ \ell\neq k}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2\frac{\pi_k\pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}.$$

In order that this estimator be unbiased, a necessary and sufficient condition is that the joint inclusion probabilities be strictly positive: $\pi_{k\ell} > 0$ for all $k, \ell \in U$ with $k \neq \ell$. Moreover a sufficient condition in order that $\widehat{\mathrm{Var}}\left(\hat{t}_{y\pi}\right) > 0$ is that $\pi_k\pi_\ell \geq \pi_{k\ell}$, for all $k, \ell \in U$, $k \neq \ell$ (YATES and GRUNDY condition, 1953). The interest of the use the unequal probabilities is that if $y_k \propto \pi_k$, then (4) equals zero. Thus, if an auxiliary variable x close to y and such that $x_k > 0,\ k \in U$, is known, it is often more interesting to select the units with unequal probabilities proportional to the x_k.

1.2. **A Problem not Really Solved**

The search for a sampling design with unequal probabilities is a relatively open problem. Indeed an infinity of solutions which respect the constraints (3) generally exists. A "good" solution should however respect the following properties:

1. The procedure should be exact in the sense that the units should be selected exactly with probabilities equalling $\pi_k,\ k \in U$.

2. The procedure should be general i.e. it should be possible to apply it to any set of first-order inclusion probabilities fixed *a priori*, which satisfy the relation (1).

3. The algorithm should be fast, the selection of the sample should be made without computing the $p(s)$ for the $N!/\{n!(N-n)!\}$ possible samples of size n.

4. The algorithm should be sequential i.e. it should be possible to apply it to a data file in only one reading by examining the units in accordance with their order number on the data file.

5. The $p(s)$ obtained by means of the sampling method should not depend on the order of the units on the data file.

6. The joint inclusion probabilities should be easy to compute without examining all the probabilities $p(s)$.

7. The joint inclusion probabilities should be strictly positive.

8. The joint inclusion probabilities should verify the Yates-Grundy condition: $\pi_{k\ell} \leq \pi_k\pi_\ell$ for all $k \neq \ell$.

9. The method should give an estimator with a smaller variance than sampling with replacement. If n units are selected independently with replacement and if the probability to select unit i is at each step $p_i = x_i / \sum_{k \in U} x_i, i \in U$, the total can be estimated by

$$\hat{t}_{yar} = \frac{1}{n} \sum_{j=1}^{n} \frac{y_j}{p_j}.$$

The variance of this estimator is

$$\mathrm{Var}(\hat{t}_{yar}) = \frac{1}{2n} \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \left(\frac{y_k}{p_k} - \frac{y_\ell}{p_\ell} \right)^2 p_k p_\ell. \tag{5}$$

By comparing expressions (4) and (5) and by considering that $\pi_k = np_k$, $k \in U$, a sufficient condition in order that a sampling without replacement be always better that sampling with replacement is that:

$$\pi_{k\ell} \geq \pi_k \pi_\ell \frac{n-1}{n}, \qquad \text{for all} \quad k, \ell \in U, \quad k \neq \ell.$$

Note that these desiderata are at the same time of a mathematical and a practical nature. Until now, a method which respects all these properties has not been discovered yet. On this subject, Deville (undated, chapter III, p. 23) pointed out that "this definition of a good unequal probability algorithm has been resisting the efforts of the statisticians for fifty years: quite a fascinating situation". It is indeed quite frustrating to notice that the most currently used exact method is the systematic sampling which is one of the oldest proposed algorithms (see MADOW, 1949, GOODMAN and KISH, 1950).

Obviously we do not claim to solve this problem once for all. In section 1, we show that some properties of sampling algorithms are contradictory. Next, in section 3, we discuss the way to build an approximation of the joint inclusion probabilities for several methods. Finally, in section 4, we present an example about the proposed approximations.

2 Remarks on the Sequential Algorithms

2.1. Joint Inclusion Probabilities

Generally, an algorithm is said to be sequential if it can be applied in only one reading of the data file. In other words, it is supposed that when the kth unit is examined, the information about the $k-1$ units which precede k is known but no information is available about the units that follow k. If the algorithm is sequential, when the decision of selecting unit k is taken,

the inclusion probabilities of the units which follow k are not known. Thus, we have the following definition.

DEFINITION 1: A sampling algorithm is said to be sequential if the probability to select any set of units $g = \{k_1, ..., k_i\} \subset U$ does not depend on the inclusion probabilities of the units which are situated after the last unit of g in the data file.

This definition is relatively restrictive. However, it corresponds to most of the algorithms where the sample is selected in only one reading of the file. Systematic sampling is obviously sequential (if the file is not randomly sorted beforehand); and so are Sunter's algorithm and Deville's method. Nevertheless, this definition excludes the algorithms where information is stocked like in the Chao procedure which selects the first n units, and subsequently updates this sample when reading the following units.

PROPOSITION 1: A general sequential unequal-probability fixed-size algorithm without replacement which provides strictly positive joint inclusion probabilities does not exist.

Proof: We shall give a counter-example of a sequential without replacement unequal probability algorithm of fixed size where a joint inclusion probability necessarily equals zero. Suppose that the algorithm is sequential, without replacement and of fixed size and take the example of an ordered population $\{1, 2, 3, 4\}$ with $n = 2$. Then, we have

$$\begin{cases} \pi_{12} + \pi_{13} + \pi_{14} = \pi_1 \\ \pi_{12} + \pi_{23} + \pi_{24} = \pi_2 \\ \pi_{13} + \pi_{23} + \pi_{34} = \pi_3 \\ \pi_{14} + \pi_{24} + \pi_{34} = \pi_4. \end{cases}$$

This system has 4 equations and 6 unknowns and can also be written

$$\begin{cases} 2(\pi_{12} - \pi_{34}) = \pi_1 + \pi_2 - \pi_3 - \pi_4 \\ 2(\pi_{13} - \pi_{24}) = \pi_1 + \pi_3 - \pi_2 - \pi_4 \\ 2(\pi_{14} - \pi_{23}) = \pi_1 + \pi_4 - \pi_2 - \pi_3 \\ \pi_{14} = \pi_1 - \pi_{12} - \pi_{13}. \end{cases} \tag{6}$$

The first equation of (6) shows that the knowledge of π_{12} implies the knowledge of π_{34}. Moreover, since the algorithm is supposed to be sequential, π_{12} does not explicitly depend either on π_3 or on π_4. However π_{12} must satisfy the following constraints:

$$\pi_{12} \leq \pi_1$$

$$\pi_{12} \leq \pi_2$$

$$\pi_{12} \geq \pi_1 + \pi_2 - 1 \tag{7}$$

$$\pi_{12} \leq \pi_1 + \pi_2 - 1 + \pi_4 \tag{8}$$

$$\pi_{12} \leq \pi_1 + \pi_2 - 1 + \pi_3. \tag{9}$$

The inequality (7) comes from (6) by considering than π_{34} must be larger than zero and that $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 2$. The inequalities (8) and (9) respectively come from the inequalities $\pi_{34} \leq \pi_3$ and $\pi_{34} \leq \pi_4$.

Since π_{12} cannot depend on π_3 and π_4, the only possible choice is

$$\pi_{12} = \begin{cases} 0 & \text{if } \pi_1 + \pi_2 \leq 1 \\ \pi_1 + \pi_2 - 1 & \text{if not.} \end{cases} \tag{10}$$

In this case, π_{34} directly comes from (6):

$$\pi_{34} = \begin{cases} 0 & \text{if } \pi_3 + \pi_4 \leq 1 \\ \pi_3 + \pi_4 - 1 & \text{if not.} \end{cases}$$

Thus we necessarily obtain a joint inclusion probability equal to zero, whether π_{12} or π_{34}. □

The proof moreover shows that if units 1 and 2 are the first two ones of the file and if the algorithm is sequential, the only possible value for π_{12} is given by (10). The Sunter Method gives here (when $n = 2$)

$$\pi_{12} = \frac{\pi_1 \pi_2}{(2 - \pi_1)}.$$

Nevertheless this algorithm is generally not exact.

2.2. **Invariance by Permutation**

The invariance by permutation is defined in the following way:

> DEFINITION 2: A sampling algorithm is said to be invariant by permutation if the probability to select a sample s does not depend on the order of the file.

Obviously, it is always possible to build a method invariant by permutation from one which is not invariant by randomly sorting the units. However, if a random sort is applied, the method is certainly no more sequential. We finally obtain the following result:

> PROPOSITION 2: A general sequential algorithm with fixed sample size without replacement and invariant by permutation generally does not exist.

Proof: By reducing to the absurd. Suppose that such an algorithm exists. By considering all the possible permutations, each couple of units $k \neq \ell$ can be situated at the beginning of the file. If we take again the particular case where $N = 4$ and $n = 2$, by (10), we have for all $k \neq \ell$

$$\pi_{k\ell} = \begin{cases} 0 & \text{if } \pi_k + \pi_\ell \leq 1 \\ \pi_k + \pi_\ell - 1 & \text{if not.} \end{cases}$$

In the case where $\pi_k = 1/2,\ k \in U$, all the $\pi_{k\ell}$ are equal to zero and thus all the $p(s)$ are equal to zero, which is impossible. Thus, a universal sequential algorithm invariant by permutation without replacement and with unequal probabilities does not exist. □

Thus a universal solution which presents all the presented properties in the previous section does not exist. It is thus not amazing that the "good" solution has not been found yet.

3 For a universal formula of the variance estimator

3.1. Systematic Sampling

Systematic sampling with unequal probabilities is a generalization of systematic sampling with equal probabilities. The method is the following: Define

$$S_k = \sum_{\ell=1}^{k} \pi_\ell, \qquad \text{for all} \quad k \in U \text{ with } S_o = 0. \tag{11}$$

Next, generate a uniform random number u in [0, 1] and select the units k such that the intervals $[S_{k-1} - u,\ S_k - u[$ include an integer. The joint inclusion probabilities for units k and ℓ only depend on the first-order inclusion probabilities and on the quantities

$$v_{k\ell} = S_{\ell-1} - S_{k-1}, \qquad k < \ell.$$

They are given by (see CONNOR, 1966, p. 388):

$$\begin{aligned} \pi_{k\ell} = &\min\{\max(0, \pi_k - \delta_{k\ell}), \pi_\ell\} \\ &+ \min\{\pi_k, \max(0, \delta_{k\ell} + \pi_\ell - 1)\}, \qquad k < \ell, \end{aligned} \tag{12}$$

where $\delta_{k\ell} = v_{k\ell} - \lfloor v_{k\ell} \rfloor$ and $\lfloor x \rfloor$ is the largest integer smaller than x.

The main problem of this method is that most of the $\pi_{k\ell}$ equal zero. In order to overcome this problem, the systematic sampling can be applied after sorting randomly the data file. The computation of the joint selection probabilities is then much more complex. Approximations of the joint inclusion probabilities can be built if an approximation of the distribution of the $v_{k\ell}$ is known. An approximation due to HARTLEY and RAO [1962] and based on an Edgeworth's development of the $v_{k\ell}$ is the following:

$$\begin{aligned} \pi_{k\ell} \approx &\frac{n-1}{n}\pi_k\pi_\ell \\ &\times \left\{ 1 + \frac{1}{n}(\pi_k + \pi_\ell) - \frac{1}{n^2}\sum_{i\in U}\pi_i^2 + \frac{2}{n^2}(\pi_k^2 + \pi_\ell^2 + \pi_k\pi_\ell) \right. \\ &\left. -\frac{3}{n^3}(\pi_k + \pi_\ell)\sum_{i\in U}\pi_i^2 + \frac{3}{n^4}(\sum_{i\in U}\pi_i^2)^2 - \frac{2}{n^3}\sum_{i\in U}\pi_i^3 \right\}. \end{aligned} \tag{13}$$

A simpler approximation was also given by DEVILLE (undated, Chapter III, p. 21) under the hypothesis that $v_{k\ell}$ has a uniform distribution $[\pi_k, n - \pi_\ell]$:

$$\pi_{k\ell} \approx \pi_k\pi_\ell \frac{n-1}{n - \pi_k - \pi_\ell}. \tag{14}$$

Deville also shows that this hypothesis of uniform distribution is verified with a large sample size.

These approximations are interesting in several respects. Indeed, one could believe that, whatever may be the sampling procedure used, if the file is first randomly sorted, the joint inclusion probabilities will be similar. Table 1 gives the square of the Euclidean distances for 3 sampling methods invariant by permutation. Each of these methods provides different joint inclusion probabilities. It is however not futile to believe that there could be possible to build a universal approximation for joint inclusion probabilities allowing to estimate the variance of the Horvitz-Thompson estimator for all sampling designs invariant by permutation.

3.2. **Adjustment on Marginal Totals**

Both proposed approximations by (13) and (14) have an annoying problem: they generally do not respect the general properties of the joint inclusion probabilities i.e.

$$\sum_{\substack{k\in U\\ k\neq\ell}} \pi_{k\ell} = \pi_\ell(n-1), \qquad \ell \in U. \tag{15}$$

Moreover these approximations do not respect the general relation:

$$\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}} \pi_{k\ell} = n(n-1). \tag{16}$$

Since

$$\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}} \pi_k\pi_\ell = n^2 - \sum_{k\in U}\pi_k^2, \tag{17}$$

another approximation which respects relation (16) can directly be proposed

$$\pi_{k\ell} \approx \pi_k\pi_\ell \frac{n(n-1)}{n^2 - \sum_{k\in U}\pi_k^2}. \tag{18}$$

An approximation which respects the constraints (15) can however be built by means of the Iterative Proportional Fitting Procedure (IPFP) (see for example BISHOP and FIENBERG [1969], and FIENBERG [1970]). First, define the matrix $\mathbf{A} = [a_{k\ell}]$ where

$$a_{k\ell} = \begin{cases} \pi_k\pi_\ell & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell. \end{cases}$$

Next, this table is adjusted by means of the IPFP method on the marginal totals defined by $b = (b_1 \ldots b_k \ldots b_N)$ and b' where $b_k = \pi_k(n-1), k \in U$. A solution for the IPFP exists if and only if there exists at least one solution respecting the constraints (15). Such a solution exists: it is given by the

systematic sampling (for example). In this case, despite the existence of zeros in $\mathbf{A}$, the existence of a solution is ensured.

In order to get symmetrical approximations of the joint inclusion probabilities at each step of the algorithm, the symmetrical version of the IPFP algorithm can be used. It can be applied in the following way: First, define the initial values for the $a_{k\ell}^{(0)}$:

$$a_{k\ell}^{(0)} = a_{k\ell}, \quad \text{for all } k, \ell. \tag{19}$$

Next, repeat the two following affectations on all the cells of the table for $i = 1, 2, 3, ...$, until a table is obtained which respects the constraints (15):

$$a_{k\ell}^{(2i-1)} = \frac{a_{k\ell}^{(2i-2)} b_k b_\ell}{\left(\sum_{k\in U} a_{k\ell}^{(2i-2)}\right)\left(\sum_{\ell\in U} a_{k\ell}^{(2i-2)}\right)},$$

$$a_{k\ell}^{(2i)} = \frac{a_{k\ell}^{(2i-1)} n(n-1)}{\sum_{k\in U}\sum_{\ell\in U} a_{k\ell}^{(2i-1)}}.$$

It can be easily shown recursively that the result is symmetrical and that it can be written

$$\tilde{\pi}_{k\ell} = \begin{cases} \beta_k \beta_\ell & \text{if } k \neq \ell \\ 0 & \text{if } k = \ell. \end{cases}$$

At the first sight, this method can look inapplicable because a table of N^2 cells must be treated. Nevertheless its application can be considerably simplified. The algorithm is initialised in the following way:

$$\beta_k^{(0)} = \pi_k, \quad \text{for all } k. \tag{20}$$

Next, the two following operations are repeated:

$$\beta_k^{(2i-1)} = \frac{(n-1)\pi_k}{(\beta^{(2i-2)} - \beta_k^{(2i-2)})}$$

$$\beta_k^{(2i)} = \beta_k^{(2i-1)} \left[\frac{n(n-1)}{(\beta^{(2i-1)})^2 - \sum_{k\in U}(\beta_k^{(2i-1)})^2}\right]^{1/2},$$

where

$$\beta^{(i)} = \sum_{k\in U} \beta_k^{(i)}, \qquad i = 1, 2, 3, ...$$

By multiplying $\beta_k^{(i)}$ and $\beta_\ell^{(i)}$, one can verify that these two algorithms are exactly the same ones. The interest of this variant is that it only needs the treatment of a vector of N elements. Moreover, at each step, the result obtained is symmetrical and respects the constraints (16). A coherent result is thus secured even if the number of iterations is limited.

It can be easily verified that the initial values given to the $\beta_k^{(0)}$ in (20) do not have any influence on the final solution. Indeed whatever may be

the strictly positive initial values given to the $\beta_k^{(0)}$, the final solution can be written as a product of two factors and is identical because of the uniqueness of the solution for the IPFP algorithm. The initial value proposed in (20) allows to shorten a little the number of steps of the algorithm because it can be expected to get β_k approximately proportional to the π_k.

Note that the approximation (18) could also be used to build initial values for the $\beta_k^{(0)}$. The Yates-Grundy estimator of the variance becomes:

$$\widehat{\mathrm{Var}}(\hat{t}_{y\pi}) \approx \sum_{k\in S} \frac{y_k^2}{\pi_k \beta_k} \sum_{\ell \in S} \frac{\pi_\ell}{\beta_\ell} - \left(\sum_{k\in S} \frac{y_k}{\beta_k} \right)^2 - n \sum_{k\in S} \frac{y_k^2}{\pi_k^2} + \hat{t}_{y\pi}^2.$$

This variance estimator can thus be got in one reading of the sample data file. One of the reviewers pointed out that if $\check{y}_k = y_k/\pi_k,\ k \in U$, denote the dilated values by the inverses of the inclusion probabilities and $\omega_k = \pi_k/\beta_k,\ k \in U$, then it is possible to write this estimator as a difference of two first-order quadratic forms:

$$\widehat{\mathrm{Var}}(\hat{t}_{y\pi}) \approx \left(\sum_{\ell \in S} \omega_\ell \right) \sum_{k\in S} \omega_k (\check{y}_k - \bar{\check{y}}_\omega)^2 - n \sum_{k\in S} \left(\check{y}_k - \frac{\hat{t}_{y\pi}}{n} \right)^2,$$

where

$$\bar{\check{y}}_\omega = \left(\sum_{k\in S} \omega_k \right)^{-1} \sum_{k\in S} \omega_k \check{y}_k.$$

4 Example

In order to measure the interest of these different approximations, we computed exactly the joint inclusion probabilities for several sampling designs. First the randomised systematic sampling is examined [System]. The joint inclusion probabilities are computed by using (12) and considering all the possible permutations. Next, Sunter's [Sunter] method is examined. This method is also applied by considering all the permutations, but it is only exact for some particular configurations (see DEVILLE and GROBRAS [1987], p. 221). An approximation of the joint inclusion probabilities is however given by Sunter:

$$\pi_{k\ell} = \frac{\pi_k \pi_\ell n(n-1)}{(n - S_{k-1})(n - S_k)} \prod_{i=1}^{k-1} \left(1 - \frac{2\pi_i}{n - S_{i-1}} \right), \qquad k < \ell.$$

Finally these inclusion probabilities are compared to the proposed approximations: the Hartley and Rao [Approx1] Approximation (13), the Deville approximation (14) [Approx2], the simple approximation method

based on the adjustment of the total [Approx3] of table **A** (3), the adjustment method to marginal totals [IPFP], the adjustment method to marginal totals by applying only one iteration [IPFP1] and two iterations [IPFP2].

Moreover the maximal entropy design [entrop] is examined. This design is the solution consisting in maximizing the quantity

$$-\sum_{\substack{s\subset U\\ \#s=n}} p(s)\log p(s),$$

under the constraints given by (3). A solution can be built by applying an algorithm of the same type as the IPFP method. This design is an interesting model because, in a certain way, it represents the most "random" design which respects the constraints (3). This design is also equivalent to the rejective Poisson sampling method of size n and to the rejective unequal probability sampling with replacement (see HÁJEK [1981], chap. 3 and 7). If the data file is randomly sorted before applying a sequential method, one could believe to obtain a result close to the maximal entropy design.

The square of the Euclidean distances between the joint inclusion probabilities were computed corresponding either to the sampling algorithms or to the proposed approximations. Since the number of iterations needed to get some solutions is large, a simple example was chosen where $N = 7$, $n = 3$ and the inclusion probabilities are the following: $\pi_1 = 0,947242$, $\pi_2 = 0,523408$, $\pi_3 = 0,504151$, $\pi_4 = 0,415621$, $\pi_5 = 0,325349$, $\pi_6 = 0,218626$ and $\pi_7 = 0,0656032$.

The squares of the Euclidean distances of the joint inclusion probabilities (multiplied by 1000) are given in Table 1.

TABLE 1

	Approx 1	Approx 2	Approx 3	IPFP	IPFP 1	IPFP 2	Entrop	Sunter	System
Approx 1	0								
Approx 2	213,89	0							
Approx 3	70,11	347,01	0						
IPFP	10,55	154,28	105,47	0					
IPFP 1	0,84	188,45	73,49	7,77	0				
IPFP 2	4,48	166,37	90,00	1,47	2,50	0			
Entrop	9,24	162,41	98,80	1,77	6,74	2,14	0		
Sunter	3,33	181,33	81,41	4,65	2,01	1,74	2,22	0	
System	11,76	158,67	104,71	1,87	9,02	3,18	1,61	4,95	0

Note that the result of the maximum entropy design is closer to the systematic design. The best approximation for systematic sampling is the adjustment on marginal totals. The results obtained by limiting to one or two iterations are valid. The simple adjustment on the general total does not seem to be a good solution. Note also the closeness of the Hartley-Rao and the IPFP approximations with one iteration. Similar results were also obtained for other examples of the same type.

5 Conclusion

The research of a "good" algorithm which satisfies the properties expressed in section 1 has resisted the statisticians' efforts for so many years because these properties are contradictory. Indeed, propositions 1 and 2 show that it is vain to search the universal sampling procedure invariant by permutation and which provides strictly positive joint inclusion probabilities. A sequential algorithm applied after a random sort can appear as an interesting solution. Since it is generally practically impossible to compute exactly the joint inclusion probabilities, one can use an approximation of these probabilities. The IPFP method allows to obtain a very coherent approximation of the joint inclusion probabilities. Moreover, the simplicity of the procedure pleads in favour of its use. Finally, it can be seen on examples that the approximations of the joint inclusion probabilities obtained with the IPFP method are much closer to the randomised systematic sampling than the other approximations generally proposed.

It is maybe possible to build a better solution by adjusting the Hartley-Rao approximation of the joint inclusion probabilities to the marginal totals. Unfortunately, such an adjustment would not allow the proposed simplification of the algorithm. Moreover, the expression of the Yates-Grundy variance could not be computed in one reading of the sample file. For these reasons, we believe that the proposed approximation is a fair compromise between precision and complexity.

• References

BISHOP, Y. M. M., FIENBERG, S.E. (1969). – "Incomplete Two-Dimensional Contingency Tables", *Biometrics*, 25, pp. 383-400.

CHAO, M. T. (1982). – "A General Purpose Unequal Probability Sampling Plan", *Biometrika*, 69, pp. 653-656.

CONNOR, W. S. (1966). – "An Exact Formula for the Probability that Specified Sampling Units will Occur in a Sample Drawn with Unequal Probabilities and without Replacement", *Journal of the American Statistical Association*, 61, pp. 384-490.

DEVILLE, J.-C. (1992). – *Une nouvelle (encore une!) méthode de tirage à probabilités inégales*, Manuscript, Paris, INSEE.

DEVILLE, J.-C. – *Cours de sondage*, Manuscript, Paris, ENSAE.

DEVILLE, J.-C. GROSBRAS J.-M. (1987). – "Algorithmes de tirage", in Droesbeke, J.-J., Fichet, B. and Tassi, P. , Eds., *Les sondages*, pp. 209-223, Paris, Economica.

FIENBERG, S. E. (1970). – "An Iterative Procedure for Estimation of Contingency Tables", *Annals of Mathematical Statistics*, 41, pp. 907-917.

GOODMAN, R., KISH, L. (1950). – "Controlled Selection – A Technique in Probability Sampling", *Journal of the American Statistical Association*, 45, pp. 350-372.

HÀJEK, J. (1981). – *Sampling from Finite Population*, New York, Marcel Dekker.

HANIF, M., BREWER, K. R. W. (1980). – "Sampling with Unequal Probabilities without Replacement: a Review", *International Statistical Review*, 48, pp. 217-335.

HARTLEY, H. O., RAO, J. N. K. (1962). – "Sampling with Unequal Probabilities and Without Replacement", *Annals of Mathematical Statistics*, 33, pp. 350-374.

MADOW, W.G. (1949). – "On the Theory of Systematic Sampling II", *Annals of Mathematical Statistics*, 20, pp. 333-354.

YATES, F, GRUNDY, P. M. (1953). – "Selection without Replacement from within Strata with Probability Proportional to Size", *Journal of the Royal Statistical Society*, 15, pp. 235-261.