Information Processing Letters 152 (2019) 105840

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl

A general result for selecting balanced unequal probability samples from a stream

Yves Tillé

University of Neuchatel, Bellevaux 51, 2000 Neuchatel, Switzerland

ARTICLE INFO

Article history:
Received 12 November 2018
Received in revised form 1 August 2019
Accepted 1 August 2019
Available online 7 August 2019
Communicated by Marek Chrobak

Keywords:
Algorithms
Balanced sampling
Chao method
Sampling
Stream

ABSTRACT

Probability sampling methods were developed in the framework of survey statistics. Recently sampling methods are the subject of a renewed interest for the reduction of the size of large data sets. A particular application is sampling from a data stream. The stream is supposed to be so huge that the data cannot be saved. When a new unit appears, the decision to conserve it or not must be taken directly without examining all the units that already appeared in the stream. In this paper, we examine the existing possible methods for sampling with unequal probabilities from a stream. Next we propose a general result about sampling in several phases from a balanced sample that enables us to propose several new solutions for sampling and multi-phase sampling from a stream. Several new applications of this general result are developed.

1. Introduction

Suppose we want to select a sample into a financial stream for the purpose of checking correctness. Very often, transactions are selected with unequal probabilities proportional to their amounts or to an index, that measures the risk of fraud. Sampling with unequal inclusion probabilities is interesting when there is a size effect, as in accounting control. In these cases, the distribution of the statistical units are often very asymmetric. Outliers can suddenly appear in the stream. Very large units contain much more information than smaller ones and must then be oversampled.

Marazzi and Tillé [13] showed the interest of using balanced sampling design. Indeed, the accuracy of the estimator is then greatly improved. If a unit is selected then a thorough control is made in order to measure a variable of interest y (for instance the amount of a fraud) that is not directly available in the stream. Our goal is to present a set of tools to be able to select such samples in very large data streams.

Sampling with unequal inclusion probabilities, fixed sample size and without replacement is more complicated than it at first appears to be. Indeed, the immediate and intuitive solutions are generally incorrect in the sense that they do not satisfy the prescribed inclusion probabilities. The first correct method was probably unequal probability systematic sampling proposed by Madow [12]. In their book, Brewer and Hanif [1] described several dozen methods of unequal probability sampling. Many of them are not general. For instance, several methods are applicable only for particular inclusion probabilities or for a sample size equal to 2. Tillé [19] presents a large set of unequal probability sampling designs. The sampling design is defined as the probability of selecting a subset of the population. A sampling algorithm is a procedure that enables to implement a sampling design. Several different algorithms can implement the same sampling design.

In this paper, we focus on one-pass (also called sequential) algorithms. For these algorithms, the decision of selecting or not a unit is taken irrevocably after examining the unit. If a unit is not selected, all the information about

E-mail address: yves.tille@unine.ch.

https://doi.org/10.1016/j.ipl.2019.105840
0020-0190/

it is forgotten. One-pass algorithms are obviously convenient to sample in a stream because the units that are not selected must not be stored. As a result, sampling methods have been gaining interest in computer science to sample in streams [6]. We give a general result that enables to define a large set of one-pass algorithms with unequal inclusion probabilities, fixed sample size and without replacement. Then, we show that a reservoir algorithm can be defined for selecting balanced samples. Very efficient methods can then be constructed without storing the information on the unselected units.

Several publications are devoted to unequal probability sampling (also called weighted sampling) from a stream [23,4,7,9,15]. Efraimidis [7] clarifies the problem by pointing out that "weighted sampling" can correspond to different definitions. The drawing probabilities are the probabilities used at each step of a procedure to select one unit at random. The inclusion probabilities are the probabilities to select a particular unit for the whole procedure. In 1953, Yates and Grundy [24] had already found that these probabilities are not equal (see also [1, Procedure 4, p. 24]). In most of the publications, "weighted sampling" is related to weighted drawing probabilities. These methods are in fact false in the sense that they do not make it possible to satisfy the fixed inclusion probabilities. We mainly focus on an extension of the Chao method [2] that enables to select a sample with fixed unequal inclusion probabilities in a stream. Chao's method exactly satisfies the fixed inclusion probabilities.

After defining the notation in Section 2, the fast cube method is briefly reminded in Section 3. Next, we explain in Section 3 how to compute and update the inclusion probabilities. When the sample size is fixed and the size of the population is not known in advance, the inclusion probabilities must be updated at each step. This is the basis of the Chao [2] method (Section 5) that uses at each step a reservoir of fixed sample size that is updated when a new unit is examined. This updating does not require the knowledge of the units that are already excluded from the sample. In Section 5, the Chao method is largely generalized. This result enables to propose several new applications in Section 6. New units can be considered by blocks, and Chao's method can be extended to balanced sampling. It is also possible to select a sample in two passes, or to select a sample in such a way that the inclusion probabilities can be changed.

2. Notation

Consider a population U_N of size N. A sample s is a subset of U_N and a sampling design $p(.)$ is a probability distribution on all the samples such that

$$p(s) \geq 0 \text{ and } \sum_{s \subseteq U_N} p(s) = 1.$$

Let also S denote a random sample that is a random variable whose distribution of probability is given by the sampling design $\Pr(S = s) = p(s)$. The size of sample S is denoted by n and can be random or not.

The inclusion probability π_k is the probability of selecting a particular unit $k \in U$. Theoretically, it can be derived from the sampling design by

$$\pi_k = \Pr(k \in S) = \sum_{\substack{s \ni k \\ s \subseteq U_N}} p(s).$$

The joint inclusion probability is the probability that two units $k, \ell \in U$ are jointly selected in the sample

$$\pi_{k\ell} = \Pr(\{k, \ell\} \subseteq S) = \sum_{\substack{s \supseteq \{k,\ell\} \\ s \subseteq U_N}} p(s).$$

In some case as in simple random sampling, stratification or maximum entropy sampling, the sampling design can be specified and a clear relation can be established between the design and the inclusion probabilities. In other cases, as in balanced sampling, the sampling design cannot be specified but the inclusion probabilities can be exactly satisfied.

A stream is a population of which the size is not known in advance. We can consider a stream as an increasing population $U_N, U_{N+1}, U_{N+2}, \ldots$ in which a sample must be selected at each step. Moreover, the population is supposed to be so large that it is not possible or difficult to save all the observations. So the sample must be selected sequentially in the stream and the sampling method ends up when the stream stops.

Consider a variable of interest y and y_k the value taken by this variable on unit k. Let Y denote the total of the values, then $Y = \sum_{k \in U_N} y_k$. If all $\pi_k > 0$ $k \in U$, then the expansion (or Horvitz-Thompson) estimator [16,10] given by

$$\widehat{Y} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

unbiasedly estimates

$$Y = \sum_{k \in U_N} y_k.$$

3. Fast cube method

Consider that a set of auxiliary variables $z_1, \ldots, z_p$ is known for each unit of the population. Let $\mathbf{z}_k \in \mathbb{R}^p$ the vector of the values taken by these p variables on unit k.

Definition 1. A sampling design with sample size n is said to be balanced for a set of auxiliary variables $z_1, \ldots, z_p$ if the Horvitz-Thompson estimators for the totals of these variables is equal to the true population totals, i.e.

$$\sum_{k \in S} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U_N} \mathbf{z}_k, \tag{1}$$

where U_N is a population of size N and S is a random sample.

Expression (1) gives the balancing equations. Balanced sampling designs are very efficient when the variable of

interest y is correlated to the auxiliary variables $z_1, \ldots, z_p$ [17,21]. An exactly balanced sample generally does not exist because the selection of a sample is an integer number problem. The aim is most often to select randomly a sample that is approximately balanced in the sense that

$$\sum_{k\in S} \frac{\mathbf{z}_k}{\pi_k} \approx \sum_{k\in U_N} \mathbf{z}_k.$$

The cube method [5,3] enables to select balanced samples. The method is composed of two phases called the *flight phase* and the *landing phase*. We will be mainly interested to the flight phase that selects quasi-samples balanced on $\mathbf{z}_k$. The details of the cube algorithm can be found in [5,3,19]. However we mainly focus on the result obtained when applying the flight phase on a vector of inclusion probabilities with a set of auxiliary variables.

Definition 2. The fast flight phase (ffph) of the cube method is a function that generates a random vector from a vector of inclusion probabilities:

$$\text{ffph}(\pi_1, \ldots, \pi_k, \ldots, \pi_N) = \boldsymbol{\psi} = (\psi_1, \ldots, \psi_k, \ldots, \psi_N)^\top$$

such that $0 \leq \psi_k \leq 1, k \in U_N$, $\mathrm{E}(\psi_k) = \pi_k$, $\text{card}\{0 < \psi_k < 1\} \leq p$ and

$$\sum_{k\in U_N} \frac{\psi_k \mathbf{z}_k}{\pi_k} = \sum_{k\in U_N} \mathbf{z}_k. \tag{2}$$

Vector $\boldsymbol{\psi}$ is a quasi-sample in the sense that almost all its components are equal to 0 or 1 except p components maximum, where p is the dimension of $\mathbf{z}_k$. Implementations are given in the fastflightcube function of the R package sampling [20] or in the flightphase function of the R package BalancedSampling [8].

4. Computation of the inclusion probabilities

Suppose that an auxiliary variable x is known for all the units of the population and let x_k be the value taken by x on unit k. The aim is to select a sample with fixed sample size by means of inclusion probabilities proportional to x_k. In a reservoir method, the first sample is composed of the first n units of the population. For each next step, an additional unit is taken into consideration and can be included with a given probability. In this case, a unit of the reservoir is removed from the sample. The inclusion probabilities must then be updated for each population size $U_n, U_{n+1}, \ldots, U_i, \ldots, U_N, \ldots$, and is computed such that

$$\pi_k(U_i, n) = \min(1, x_k\, \tau_i), \tag{3}$$

where τ_i is obtained by solving

$$\sum_{k\in U_i} \min(1, x_k\, \tau_i) = n.$$

However, Cohen et al. [4] have shown that a sample can be selected without computing all the $\pi_k(U_i, n)$ at each step.

5. Chao reservoir method

For equal inclusion probabilities a "reservoir" method is described in Knuth [11, p. 144], McLeod and Bellhouse [14] and Vitter [22]. The only reservoir method that really satisfies given inclusion probabilities was proposed by Chao [2] [see also18]. Consider a sequence of population $U_n \subsetneq U_{n+1} \subsetneq \cdots \subsetneq U_i \subsetneq \cdots \subsetneq U_N$. Let $\pi_k(U_i, n)$ denote the inclusion probabilities computed on a population U_i of size i for a sample of size n as defined in Equation (3). Chao's method begins by selecting the first n population units U_n, which is called the reservoir. At each step, the reservoir is updated as follows. Suppose that at step $i-1$, the reservoir is denoted by S_{i-1}. At step $i = n+1, \ldots, N$, unit i is included in the reservoir with a probability $\pi_i(U_i, n)$. If unit i is selected, one of the units of the reservoir is removed with probability

$$a_{ki} = \frac{1}{\pi_k(U_i, n)} \left[1 - \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}\right], k = 1, \ldots, i-1.$$

It is indeed possible to prove that

$$\sum_{k\in S_{i-1}} a_{ki} = 1.$$

Another way of presenting the method consists of adding unit i in the reservoir that becomes $A_i = S_{i-1} \cup \{i\}$. Next n units are selected from A_i with probabilities

$$\frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}, k \in S_{i-1}, \text{ and } \pi_i(U_i, n).$$

Indeed,

$$\sum_{s\in S_{i-1}} \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)} + \pi_i(U_i, n) = n. \tag{4}$$

In the original Chao paper, it is assumed that the inclusion probabilities $\pi_k(U_i, n)$ are computed from U_i before implementing the method. The $\pi_k(U_i, n)$ are then recomputed at each step for all the units of U_i. However, Cohen et al. [4] have shown that the unselected units can be forgotten and that a reservoir implementation is possible in the sense where only the units selected at step i must be stored. Indeed, it is possible to compute $\pi_k(U_i, n)$ by knowing only the values x_k of S_{i-1}. This shortcut is a very important improvement of the method, because in a very large data stream, it is not necessary to store the information about unselected units.

6. Generalization of Chao's method

The results of Chao [2] and Cohen et al. [4] can be generalized in two directions. The same principle can be used to select balanced samples and not only samples with fixed sample size. More complex algorithm in two passes or by blocks can also be implemented. Suppose that a quasi-sample $\boldsymbol{\psi}^1 = (\psi_1^1, \ldots, \psi_N^1)^\top$ has been selected with balanced sampling from a population U_N with inclusion probabilities $\boldsymbol{\pi}^1 = (\pi_1^1, \ldots, \pi_N^1)^\top$. The inclusion probabil-

ities are proportional to a positive variable x, so we can write $\pi_k^1 = \min(1, \tau_1 x_k)$, where τ_1 is obtained by solving

$$\sum_{k\in U_N} \min(1, \tau_1 x_k) = n.$$

The theory below admits non-integer values for n.

The quasi-sample $\boldsymbol{\psi}^1$ is supposed to be balanced on p auxiliary variables whose values for unit k are the components of vector $\mathbf{z}_k$. The balancing equations are then exactly satisfied

$$\sum_{k\in U_N} \frac{\mathbf{z}_k \psi_k^1}{\pi_k^1} = \sum_{k\in U_N} \mathbf{z}_k. \tag{5}$$

Proposition 1. *If there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^\top \mathbf{z}_k = x_k$, for all $k \in \{k \in U_n | \pi_k^1 < 1\}$, Equation (5) implies: $\sum_{k\in U_N} \psi_k^1 = \sum_{k\in U_N} \pi_k^1$.*

Proof. From left side of Equation (5),

$$\begin{aligned}
\boldsymbol{\lambda}^\top \sum_{k\in U_N} \frac{\mathbf{z}_k \psi_k^1}{\pi_k^1} &= \sum_{k\in U_N} \frac{x_k \psi_k^1}{\pi_k^1} \\
&= \sum_{k\in U_N|\pi_k^1<1} \frac{x_k \psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \frac{x_k \psi_k^1}{\pi_k^1} \\
&= \sum_{k\in U_N|\pi_k^1<1} \frac{x_k \psi_k^1}{x_k \tau_1} + \sum_{k\in U_N|\pi_k^1=1} x_k \\
&= \sum_{k\in U_N|\pi_k^1<1} \frac{\psi_k^1}{\tau_1} + \sum_{k\in U_N|\pi_k^1=1} x_k. \qquad (6)
\end{aligned}$$

From right side of Equation (5),

$$\boldsymbol{\lambda}^\top \sum_{k\in U_N} \mathbf{z}_k = \sum_{k\in U_N} x_k. \tag{7}$$

By equating (6) and (7), we obtain

$$\sum_{k\in U_N|\pi_k^1<1} \psi_k^1 = \sum_{k\in U_N|\pi_k^1<1} \tau_1 x_k = \sum_{k\in U_N|\pi_k^1<1} \pi_k^1.$$

When $\pi_k^1 = 1$, $\psi_k^1 = 1$, and we obtain Proposition 1. □

At the second phase, another balanced quasi-sample ψ_k^2 if selected in such a way that the final inclusion probabilities are $\boldsymbol{\pi}^2 = (\pi_1^2, \ldots, \pi_k^2 \ldots, \pi_N^2)^\top$. The inclusion probabilities π_k^2 are supposed to be proportional to another variable v_k, i.e. $\pi_k^2 = \min(1, \tau_2 v_k)$ where τ_2 is obtained by solving

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) = m.$$

Variable v_k can be equal to x_k. The theory below also admits non-integer values for m. Moreover the components of $\boldsymbol{\pi}^2$ must all be less or equal to the components of $\boldsymbol{\pi}^1$ ($\pi_k^2 \le \pi_k^1, k \in U_N$), and thus $m \le n$.

The second phase sample is drawn in the quasi-sample $\boldsymbol{\psi}^1$. The drawing probabilities are:

$$\xi_k = \frac{\pi_k^2 \psi_k^1}{\pi_k^1}. \tag{8}$$

Since $\mathrm{E}(\psi_k^1) = \pi_k^1$, we have $\mathrm{E}(\xi_k) = \pi_k^2$.

Proposition 2. *If the quasi-sample ψ_1 is balanced on $\mathbf{z}_k$ and there exists a vector $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\boldsymbol{\theta}^\top \mathbf{z}_k = v_k$, then the inclusion probabilities π_k^2 and thus the drawing probabilities ξ_k can be computed from v_k without knowing the units such that $\psi_k^1 = 0$, by solving in τ_2:*

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) \frac{\psi_k^1}{\pi_k^1} = m.$$

Proof. From the balancing equations, we have

$$\sum_{k\in U_N} v_k \frac{\psi_k^1}{\pi_k^1} = \sum_{k\in U_N} v_k.$$

Moreover, if $\pi_k^1 < 1$, we also have $\pi_k^2 < 1$. Consider

$$\begin{aligned}
\sum_{k\in U_N} \pi_k^2 \frac{\psi_k^1}{\pi_k^1} &= \sum_{k\in U_N|\pi_k^1<1} \pi_k^2 \frac{\psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N|\pi_k^1<1} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} - \sum_{k\in U_N|\pi_k^1=1} \tau_2 v_k \frac{\psi_k^1}{\pi_k^1} \\
&\quad + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \tau_2 v_k - \sum_{k\in U_N|\pi_k^1=1} \tau_2 v_k + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N|\pi_k^1<1} \pi_k^2 + \sum_{k\in U_N|\pi_k^1=1} \pi_k^2 \\
&= \sum_{k\in U_N} \pi_k^2.
\end{aligned}$$

So we obtain

$$\sum_{k\in U_N} \min(1, \tau_2 v_k) \frac{\psi_k^1}{\pi_k^1} = \sum_{k\in U_N} \min(1, \tau_2 v_k). \quad \square$$

Proposition 2 enables us to forget the units that are not selected while preserving the possibility of computing the second-phase inclusion probabilities. The second quasi-sample is balanced on $\psi_k^1 \mathbf{z}_k / \pi_k^1$. Thus

$$\sum_{k\in U_N} \frac{\mathbf{z}_k \psi_k^2}{\pi_k^2} = \sum_{k\in U_N} \frac{\mathbf{z}_k \psi_k^1 \psi_k^2}{\xi_k \pi_k^1} = \sum_{k\in U_N} \frac{\mathbf{z}_k \psi_k^1}{\pi_k^1} = \sum_{k\in U_N} \mathbf{z}_k. \tag{9}$$

So, if $\sum_{k\in U_N} \pi_k^1$ is integer, $\boldsymbol{\psi}^1$ is a sample with fixed sample size. Proposition 2 also implies that if there exists a

vector $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}^\top \mathbf{z}_k = v_k$, for all $k \in U_n$, then

$$\sum_{k\in U_N} \xi_k = \sum_{k\in U_N} \pi_k^2. \tag{10}$$

Equation (4) for the Chao sampling design is a particular case of equality (10) when $x_k = v_k$ and $\mathbf{z}_k = x_k$.

Moreover, if the second quasi-sample is balanced on $\psi_k^1 \mathbf{z}_k/\pi_k^1 = \psi_k^1$, we also have:

$$\sum_{k\in U_N|\psi_k^1>0} \frac{\psi_k^1\psi_k^2}{\xi_k} = \sum_{k\in U_N} \frac{\pi_k^1\psi_k^2}{\pi_k^2} = \sum_{k\in U_N} \psi_k^1 = \sum_{k\in U_N} \pi_k^1.$$

7. Some applications

With the result of the previous section, it is possible to imagine several new methods.

7.1. Preserving two variables for unequal inclusion probability sampling

Suppose that two variables $u_k > 0$ and $v_k > 0$ are available and that one hesitates on which one the inclusion probabilities should be computed. For example, we want to select transactions in a financial stream and we do not know in advance if it is better to select the units with equal probabilities or with probabilities proportional to transaction amounts. The trick consists of selecting a first sample that preserve the possibilities of selecting a sample in a second phase that can be either with inclusion probabilities proportional to either u_k or v_k.

First compute $x_k = \max(u_k, v_k)$. A first quasi-sample ψ_k^1 can be selected with inclusion probabilities $\pi_k^1 = \min(1, \tau_1 x_k)$, where τ_1 can be chosen freely. This sample is balanced on three auxiliary variables $\mathbf{z}_k = (u_k, v_k, x_k)^\top$. Next a second sample can be selected with inclusion probabilities $\pi_k^2 = \min(1, \tau_2 u_k)$ or $\pi_k^2 = \min(1, \tau_2 v_k)$ with $\tau_2 \leq \tau_1$.

7.2. Block reservoir method

The Chao method can be generalized when blocks of units appear together in the stream. In this case, it is possible to treat the whole block at once. In place of considering at each step one unit to be included in the reservoir, a set of H units can be considered (see on this topic [4]). At the first step, a block on the first H units (where $H > n$) are taken in the population with inclusion probabilities $\pi_k^1 = \min(1, \tau_1 x_k)$, where

$$\sum_{k\in U_H} \min(1, \tau_1 x_k) = n.$$

From these H units n are selected in a sample or a quasi-sample ψ_k^1. Next, the H following units are considered, and for the $n + H$ units, with inclusion probabilities $\pi_k^2 = \min(1, \tau_2 x_k)$, where

$$\sum_{k\in U_{2H}} \min(1, \tau_2 x_k) = \sum_{k\in U_{2H}} \min(1, \tau_2 x_k) \frac{\psi_k^1}{\pi_k^1} = n.$$

Drawing probabilities are computed using Equation (8) and next n units are selected from these $n + H$ units, and so on.

7.3. Balanced reservoir method

The method of Chao can be generalized to balanced sampling. This generalization enables to select directly a balanced sample from a stream as for example from a stream of financial transactions. Consider that p balancing variables are in $\mathbf{z}_k$ and that there exists a vector $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^\top \mathbf{z}_k = x_k$. Take $n + 1$ units of the population with inclusion probabilities proportional to x_k. Then $\pi_k^1 = \min(1, \tau_1 x_k)$ where

$$\sum_{k\in U_{n+1}} \min(1, \tau_1 x_k) = n.$$

Select a quasi-sample ψ_k^1 that is balanced with size n in this subset of the population. Since $\boldsymbol{\lambda}^\top \mathbf{z}_k = x_k$, then $\sum_{k\in U_{n+1}} \pi_k^1 = n$.

Take the next unit (number $n + 2$). Compute $\pi_k^2 = \min(1, \tau_2 x_k)$ where

$$\sum_{k\in U_{n+2}} \min(1, \tau_2 x_k) = n.$$

Select a balanced quasi-sample with drawing probabilities given by Equation (8), and so on. This method is quite slow, because a balanced sample must be selected $N - n$ times.

7.4. Balanced block reservoir method

The block reservoir method can be combined with balanced sampling. The method is more efficient than the previous one because all the units of each block are treated together. A block on the first H units (where $H > n$) are taken in the population. From these H units a quasi-sample is selected using balanced sampling. The next H units are taken, inclusion probabilities are computed using Equation (8). Balanced sampling is applied again to select a quasi-sample, and so on. This method is faster than the previous one because the flight phase of the cube method must be run less frequently.

7.5. Two-pass method

In case of a very large stream with a sample size that is also very large, one can imagine a two-pass procedure. In the first pass, the size of the stream is considerably reduced. A second pass is needed to select a sample. The advantage of this method is that the inclusion probabilities are very easy to calculate, which makes the first pass very fast.

During the first phase, the first n units are selected. Next a balanced quasi-sample is selected with inclusion probabilities $\pi_k^1 = \pi_k(U_k, n), k = n + 1, \ldots, N$. $\pi_k(U_i, n) = \min(\tau_i x_k, 1)$ where

$$\sum_{k\in U_i} \min(\tau_i x_k, 1) = n$$

This quasi-sample ψ_k^1 is balanced on two auxiliary variables $\mathbf{z}_k = (\pi_k(U_k, n), x_k)^\top$.

If x_k is constant, the $\pi_k(U_k, n) = 1, k = 1, \ldots, n$, and $\pi_k(U_k, n) = n/k, k = n+1, \ldots, N$. In this case the expected sample size is

$$\sum_{k \in U_N} \pi_k(U_k, n) = n + \sum_{k=n+1}^{N} \frac{n}{k} = n\left(1 + \sum_{k=1}^{N} \frac{1}{k} - \sum_{k=1}^{n} \frac{1}{k}\right) \approx n(1 + \ln N - \ln n) = n + n \ln \frac{N}{n}.$$

The sample size is considerably reduced at the end of the first phase.

Next, in a second phase, a sample is selected with inclusion probabilities $\pi_k^2 = \pi_k(U_N, n)$ for all $k \in U_N$. The drawing probabilities are given by Equation (8). Note that $\pi_k(U_k, n)$ should be computed at each step that can be sometimes slow. However, any an upper bound for $\pi_k(U_k, n)$ can also be used. For instance, if

$$\pi_k(U_k, n) = \min(x_k \tau_k, 1),$$

then an upper bound easily computable could be

$$\pi_{k+1}(U_{k+1}, n) \leq \min\left(n x_k \frac{\sum_{k \in U_k} x_k}{\sum_{k \in U_{k+1}} x_k}, 1\right).$$

8. Discussion

The Chao method enables to sample with unequal inclusion probabilities from a stream because the inclusion probabilities can be updated without knowing the units that are not selected. This result can be generalized for quasi-samples obtained by the cube method. Moreover, these results can also be extended for problems where blocks of units appears in the stream. Therefore, one can define several new algorithms to sample from a stream while preserving balancing properties of the sample.

Declaration of competing interest

I wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

I confirm that the manuscript has been read and approved by the only author and that there are no other persons who satisfied the criteria for authorship but are not listed.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

I further confirm this manuscript that does not involve either experimental animals or human patients.

I understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office).

I confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from (yves.tille@unine.ch).

Acknowledgements

The author gratefully thanks the two reviewers for their constructive comments and recommendations, which certainly help to improve the readability and quality of the paper.

References

[1] K.R.W. Brewer, M. Hanif, Sampling with Unequal Probabilities, Springer, New York, 1983.
[2] M.T. Chao, A general purpose unequal probability sampling plan, Biometrika 69 (1982) 653–656.
[3] G. Chauvet, Y. Tillé, A fast algorithm of balanced sampling, Comput. Stat. 21 (2006) 9–31.
[4] E. Cohen, N. Duffield, H. Kaplan, C. Lund, M. Thorup, Stream sampling for variance-optimal estimation of subset sums, in: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, 2009, pp. 1255–1264.
[5] J.C. Deville, Y. Tillé, Efficient balanced sampling: the cube method, Biometrika 91 (2004) 893–912.
[6] N. Duffield, Sampling for passive internet measurement: a review, Stat. Sci. (2004) 472–498.
[7] P.S. Efraimidis, Weighted random sampling over data streams, in: C. Zaroliagis, G. Pantziou, S. Kontogiannis (Eds.), Algorithms, Probability, Networks, and Games: Scientific Papers and Essays Dedicated to Paul G. Spirakis on the Occasion of His 60th Birthday, Springer International Publishing, Cham, 2015, pp. 183–195.
[8] A. Grafström, J. Lisic, BalancedSampling: balanced and spatially balanced sampling. R package version 1.5.2, 2016.
[9] P.J. Haas, Data-stream sampling: basic techniques and results, in: M. Garofalakis, J. Gehrke, R. Rastogi (Eds.), Data Stream Management, Springer, 2016, pp. 13–44.
[10] D.G. Horvitz, D.J. Thompson, A generalization of sampling without replacement from a finite universe, J. Am. Stat. Assoc. 47 (1952) 663–685.
[11] D.E. Knuth, The Art of Computer Programming (Volume II): Seminumerical Algorithms, Addison-Wesley, Reading, MA, 1981.
[12] W.G. Madow, On the theory of systematic sampling, II, Ann. Math. Stat. 20 (1949) 333–354.
[13] A. Marazzi, Y. Tillé, Using past experience to optimize audit sampling design, Rev. Quant. Finance Account. 49 (2017) 435–462.
[14] A.I. McLeod, D.R. Bellhouse, A convenient algorithm for drawing a simple random sampling, Appl. Stat. 32 (1983) 182–184.
[15] K. Müller, Accelerating weighted random sampling without replacement, Arbeitsberichte Verkehrs- und Raumplanung 1141, 2016.
[16] R.D. Narain, On sampling without replacement with varying probabilities, J. Indian Soc. Agric. Stat. 3 (1951) 169–174.
[17] D. Nedyalkova, Y. Tillé, Optimal sampling and estimation strategies under linear model, Biometrika 95 (2008) 521–537.
[18] R.A. Sugden, T.M.F. Smith, R.P. Brown, Chao's list sequential scheme for unequal probability sampling, J. Appl. Stat. 23 (1996) 413–421.
[19] Y. Tillé, Sampling Algorithms, Springer, New York, 2006.
[20] Y. Tillé, A. Matei, Sampling: survey sampling. R package version 2.8, 2016.
[21] Y. Tillé, M. Wilhelm, Probability sampling designs: balancing and principles for choice of design, Stat. Sci. 32 (2017) 176–189.
[22] J.S. Vitter, Random sampling with a reservoir, ACM Trans. Math. Softw. 11 (1985) 37–57.
[23] C.K. Wong, M.C. Easton, An efficient method for weighted sampling without replacement, SIAM J. Comput. 9 (1980) 111–113.
[24] F. Yates, P.M. Grundy, Selection without replacement from within strata with probability proportional to size, J. R. Stat. Soc. B 15 (1953) 235–261.