Published in Handbook of Statistics 29, part a, 39-54, 2009 which should be used for any reference to this work

# Sampling with Unequal Probabilities

*Yves G. Berger*[1] *and Yves Tillé*[2]

[1]*Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK*
[2]*Institute of Statistics, University of Neuchâtel, Switzerland*

## 1. Introduction

Since the mid 1950s, there has been a well-developed theory of sample survey design inference embracing complex designs with stratification and unequal probabilities (Smith, 2001). Unequal probability sampling was first suggested by Hansen and Hurwitz (1943) in the context of sampling with replacement. Narain (1951), Horvitz and Thompson (1952) developed the corresponding theory for sampling without replacement. A large part of survey sampling literature is devoted to unequal probabilities sampling, and more than 50 sampling algorithms have been proposed. Two books (Brewer and Hanif, 1983; Tillé, 2006) provide a summary of these methods.

Consider a finite population $U$ of size $N$. Each unit of the population can be identified by a label $k = 1, \ldots, N$. A sample $s$ is a subset of $U$. A sampling design $p(.)$ is a probability measure on all the possible samples so that

$$p(s) \geq 0, \text{ for all } s \in U, \text{ and } \sum_{s \in U} p(s) = 1.$$

Let $n(s)$ denote the size of the sample $s$. When the sample size is not random, we denote the sample size by $n$. An unequal probability sampling design is often characterized by its first-order inclusion probabilities given by $\pi_k = p(k \in s)$. The joint inclusion probabilities of unit $k$ and $\ell$ are defined by $\pi_{k\ell} = p(k \in s \text{ and } \ell \in s)$.

Suppose we wish to estimate the population total

$$Y = \sum_{k \in U} y_k$$

of a characteristic of interest $y$, where $y_k$ is the value of a unit labeled $k$. An estimator of $Y$ is given by the $\pi$-estimator (Horvitz and Thompson, 1952; Narain, 1951) defined by

$$\widehat{Y}_\pi = \sum_{k \in s} \frac{y_k}{\pi_k}.$$

This estimator is design unbiased provided that all the $\pi_k > 0$.

Under unequal probability sampling, the variance of $\widehat{Y}_\pi$ may be considerably smaller than the variance under an equal probability sampling design (Cochran, 1963), when the correlation between the characteristic of interest and the first-order inclusion probabilities is strong. Alternative estimators when this correlation is weak are discussed in Section 3.

It is common practice to use inclusion probabilities that are proportional to a known positive size variable $x$. In this case, the inclusion probabilities are computed as follows

$$\pi_k = \frac{nx_k}{X}, \tag{1}$$

where $X = \sum_{k\in U} x_k$, assuming $nx_k \leq X$ for all $k$. If $nx_k > X$, we set $\pi_k = 1$ and we recalculate the $\pi_k$ using (1) on the remaining units after substituting $n$ with $n$ subtracted by the number of $\pi_k$ equal to 1.

Another application of unequal probability sampling design is with multistage sampling, where the selection of primary units within strata may be done with unequal probability. For example, self-weighted two-stage sampling is often used to select primary units with probabilities that are proportional to the number of secondary units within the primary units. A simple random sample is selected within each primary unit.

The variance of the $\pi$-estimator plays an important role in variance estimation, as most estimators of interest can be linearized to involve $\pi$-estimators (see Section 5). The sampling variance of $\widehat{Y}_\pi$ is given by

$$\text{var}\,(\widehat{Y}_\pi) = \sum_{k\in U}\sum_{\ell\in U}(\pi_{k\ell} - \pi_k\pi_\ell)\frac{y_k y_\ell}{\pi_k\pi_\ell}.$$

Horvitz and Thompson (1952) proposed an unbiased estimator of var $(\widehat{Y}_\pi)$:

$$\text{var}\,(\widehat{Y}_\pi) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}\frac{y_k y_\ell}{\pi_k\pi_\ell}. \tag{2}$$

If the sample size is fixed, Sen (1953), Yates and Grundy (1953) proposed another estimator of var $(\widehat{Y}_\pi)$:

$$\widehat{\text{var}}\,(\widehat{Y}_\pi) = \frac{1}{2}\sum_{k\in s}\sum_{\ell\in s}\frac{\pi_k\pi_\ell - \pi_{k\ell}}{\pi_{k\ell}}\left(\frac{y_k}{\pi_k} - \frac{y_\ell}{\pi_\ell}\right)^2. \tag{3}$$

This estimator is design unbiased when $\pi_{k\ell} > 0$ for all $k, \ell \in U$. It can take negative values unless $\pi_k\pi_\ell - \pi_{k\ell} \geq 0$, $k \neq \ell \in U$. However, it is rarely used because the joint inclusion probabilities are sometimes difficult to compute and because the double sum makes (3) computationally intensive. In Section 4, we show that, in particular cases, the variance can be estimated without joint inclusion probabilities.

## 2. Some methods of unequal probability sampling

### *2.1. Poisson sampling*

Poisson sampling was proposed by Hájek (1964) and discussed among others in Ogus and Clark (1971), Brewer et al. (1972, 1984), and Cassel et al. (1993, p. 17). Each unit

of the population is selected independently with a probability $\pi_k$. The sample size $n(s)$ is therefore random. All the samples $s \subset U$ have a positive probability of being selected and there is thus a non-null probability of selecting an empty sample. The sampling design is given by

$$P(s) = \left[\prod_{k\in s} \frac{\pi_k}{1-\pi_k}\right]\left[\prod_{k\in U}(1-\pi_k)\right], \text{ for all } s \subset U.$$

Since the units are selected independently, we have that $\pi_{k\ell} = \pi_k\pi_\ell$, for all $k \neq \ell$.

The variance of the $\pi$-estimator, given in (2), reduces to

$$\text{var}\left(\widehat{Y}_\pi\right) = \sum_{k\in U} \frac{1}{\pi_k}(1-\pi_k)y_k^2,$$

which can be unbiasedly estimated by

$$\widehat{\text{var}}\left(\widehat{Y}_\pi\right) = \sum_{k\in s}(1-\pi_k)\frac{y_k^2}{\pi_k^2}.$$

The estimator of variance is simple because it does not involve joint inclusion probabilities. Note that the Poisson sampling design maximizes the entropy (Hájek, 1981, p.29) given by

$$I(p) = -\sum_{s\subset U} p(s)\log p(s), \tag{4}$$

subject to given inclusion probabilities $\pi_k, k \in U$. Since the entropy is a measure of randomness, the Poisson sampling design can be viewed as the most random sampling design that satisfies given inclusion probabilities.

Poisson sampling is rarely applied in practice because its sample size is random implying a nonfixed cost of sampling. This design is, however, often used to model nonresponse. Moreover, Poisson sampling will be used in Section 2.7 to define the conditional Poisson sampling design. This sampling design is also called the maximum entropy design with fixed sample size. The use of design that maximizes the entropy is useful because it allows a simple estimation for the variance.

### *2.2. Sampling with replacement*

Unequal probability sampling with replacement is originally due to Hanssen and Hurwitz. Properties of this design are widely covered in the literature (Bol'shev, 1965; Brown and Bromberg, 1984; Dagpunar, 1988; Davis, 1993; Devroye, 1986; Ho et al., 1979; Johnson et al., 1997; Kemp and Kemp, 1987; Loukas and Kemp, 1983; Tillé, 2006).

Consider selection probabilities $p_k$ that are proportional to a positive size variable $x_k, k \in U$; that is,

$$p_k = \frac{x_k}{\sum_{\ell\in U} x_\ell}, k \in U.$$

A simple method to select a sample with unequal probabilities with replacement consists in generating a uniform random number $u$ in $[0, 1]$ and selecting unit $k$ so that AQ:1

$v_{k-1} \leq u < v_k$, where

$$v_k = \sum_{\ell=1}^{k} p_\ell, \text{ with } v_0 = 0.$$

This process is repeated independently $m$ times. Note that there are more efficient algorithms that may be used to select a sample with replacement with unequal probabilities (Tillé, 2006, p. 75).

Let $\tilde{y}_i$ denote the value of the characteristic of the $i$th selected unit and $\tilde{p}_i$, its associated selection probability. Note that, under sampling with replacement, the same unit can be selected several times. The ratios $\tilde{y}_i/\tilde{p}_i$ are $n$ independent random variables. The total $Y$ can be estimated by the Hansen–Hurwitz estimator

$$\widehat{Y}_{\text{HH}} = \frac{1}{m} \sum_{i=1}^{m} \frac{\tilde{y}_i}{\tilde{p}_i}.$$

This estimator is design unbiased as

$$\text{E}\left(\widehat{Y}_{\text{HH}}\right) = \frac{1}{m} \sum_{i=1}^{m} \text{E}\left(\frac{\tilde{y}_i}{\tilde{p}_i}\right) = \frac{1}{m} \sum_{i=1}^{m} Y = Y.$$

The variance of $\widehat{Y}_{\text{HH}}$ is given by

$$\text{var}\left(\widehat{Y}_{\text{HH}}\right) = \frac{1}{m} \sum_{k \in U} p_k \left(\frac{y_k}{p_k} - Y\right)^2,$$

which can be unbiasedly estimated by

$$\widehat{\text{var}}\left(\widehat{Y}_{\text{HH}}\right) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{\tilde{y}_i}{\tilde{p}_i} - \widehat{Y}_{\text{HH}}\right)^2. \tag{5}$$

The Hansen–Hurwitz estimator is not the best estimator as it is not admissible because it depends on the multiplicity of the units (Basu, 1958, 1969; Basu and Ghosh, 1967). Nevertheless, the Hansen–Hurwitz variance estimator can be used to approximate the variance of the Horvitz–Thompson estimator under sampling without replacement when $m/N$ is small.

Sampling without replacement may lead to a reduction of the variance compared to sampling with replacement (Gabler, 1981, 1984). A design without replacement with inclusion probabilities $\pi_k$ is considered to be a good design if the Horvitz–Thompson estimator is always more accurate than the Hansen–Hurwitz estimator under sampling with replacement with probabilities $p_k = \pi_k/n$. Gabler (1981, 1984) gave a condition under which this condition holds. For example, this condition holds for the Rao–Sampford design given in Section 2.4 and for the maximum entropy design with fixed sample size (Qualité, 2008).

### *2.3. Systematic sampling*

Systematic sampling is widely used by statistical offices due to its simplicity and efficiency (Bellhouse, 1988; Bellhouse and Rao, 1975; Berger, 2003; Iachan, 1982, 1983).

This sampling design has been studied since the early years of survey sampling (Cochran, 1946; Madow, 1949; Madow and Madow, 1944). There are two types of systematic sampling: a systematic sample can be selected from a deliberately ordered population or the population can be randomized before selecting a systematic sample. The latter is often called randomized systematic design.

In many practical situations, it is common practice to let the population frame have a predetermined order. For example, a population frame can be sorted by a size variable, by region, by socioeconomic group, by postal sector, or in some other way. In this case, systematic sampling is an efficient method of sampling (Iachan, 1982). Systematic sampling from a deliberately ordered population is generally more accurate than randomized systematic sampling (Särndal et al., 1992, p. 81), especially when there is a trend in the survey variable $y$ (Bellhouse and Rao, 1975).

A systematic sample is selected as follows. Let $u$ be a random number between 0 and 1 generated from a uniform distribution. A systematic sample is a set of $n$ units labeled $k_1, k_2, \ldots, k_n$ such that $\pi^{(c)}_{k_\ell - 1} < u + \ell - 1 \leq \pi^{(c)}_{k_\ell}$, where $\ell = 1, \ldots, n$ and

$$\pi^{(c)}_k = \sum_{\substack{j \in U \\ j \leq k}} \pi_j.$$

In the special case where $\pi_k = n/N$, this design reduces to the customary systematic sampling design, where every $a$th unit is selected and $a = \lfloor N/n \rfloor$.

The systematic design with a deliberately ordered population suffers from a serious flaw, namely, that it is impossible to unbiasedly estimate the sampling variance (Iachan, 1982), and customary variance estimators given in (3) are inadequate and can overestimate the variance significantly (Särndal et al., 1992, Chapter 3).

Systematic sampling from a randomly ordered population consists in randomly arranging the units, giving the same probability to each permutation, since random ordering is part of the sampling design. This design was first suggested by Madow (1949). Hartley and Rao (1962) developed the corresponding asymptotic theory for large $N$ and small sampling fraction. Under randomized systematic sampling, Hartley and Rao (1962) derived a design unbiased variance estimator (see Section 4).

For the randomized systematic design, the joint inclusion probabilities are typically positive and the variance can be unbiasedly estimated (Hájek, 1981; Hartley and Rao, 1962). With a deliberately ordered population, alternative estimators for the variance can be used (Bartolucci and Montanari, 2006; Berger, 2005a; Brewer, 2002, Chapter 9).

### *2.4. Rao–Sampford sampling design*

The Rao–Sampford sampling design (Rao, 1965; Sampford, 1967) is a popular design used for unequal probability sampling without replacement. It is implemented by selecting the first unit with drawing probabilities $p_k = \pi_k/n$. The remaining $n - 1$ units are selected with replacement with drawing probabilities that are proportional to $\pi_k/(\pi_k - 1)$. The sample is accepted if the $n$ units drawn are all distinct, otherwise, it is rejected and the process is repeated. The first-order inclusion probabilities are exactly given by $\pi_k$. Sampford (1967) derived an exact expression for the joint inclusion probabilities $\pi_{k\ell}$.

The main advantage of this design is its simplicity. It also has a simple expression for the variance (see Section 4). However, this design is not suitable when the $\pi_k$ are large,

as we would almost surely draw the units with large $\pi_k$ at least twice and it would not be possible to select one Rao–Sampford sample. For example, consider $N = 86$, $n = 36$, and $\pi_k$ proportional to $(k/100)^5 + 1/5$. The probability that all the units drawn from subsequent independent draws will be distinct is approximately $10^{-36}$ (Hájek, 1981, p. 70), which is negligible. Nevertheless, Tillé (2006, p. 136) and Bondesson et al. (2006) suggested several alternative algorithms to implement the Rao–Sampford design.

### *2.5. Sampling by the splitting method*

The splitting method, proposed by Deville and Tillé (1998), is a general class of sampling without replacement with fixed sample size and unequal probabilities. First, each inclusion probability is split into two or more quantities. Secondly, one of these sets of quantities is randomly selected in such a way that the overall inclusion probabilities are equal to $\pi_k$. These steps are repeated until a sample is obtained.

This method can be implemented as follows. First, $\pi_k$ is split into two quantities $\pi_k^{(1)}$ and $\pi_k^{(2)}$, which satisfy the following relation:

$$\pi_k = \lambda\pi_k^{(1)} + (1-\lambda)\pi_k^{(2)},$$

with

$$0 \le \pi_k^{(1)} \le 1 \text{ and } 0 \le \pi_k^{(2)} \le 1,$$

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

where $\lambda$ is any constant such that $0 < \lambda < 1$.

The method consists of choosing

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

After this first step, any design can be used to select a sample with inclusion probabilities $\pi_k^{(1)}$ or $\pi_k^{(2)}$. If some of the $\pi_k^{(1)}$ or $\pi_k^{(2)}$ are all equal to 0 or 1, we would sample from a smaller population. The splitting can in turn be used to select a sample with probabilities $\pi_k^{(1)}$ or $\pi_k^{(2)}$. We could also choose $\pi_k^{(1)}$ in such a way that the $\pi_k^{(1)}$ are all equal. In this case, simple random sampling without replacement can be used.

This approach can be generalized to a splitting method into $M$ sets of inclusion probabilities. First, we construct the $\pi_k^{(j)}$ and the $\lambda_j$ in such a way that

$$\sum_{j=1}^{M} \lambda_j = 1,$$

where

$$0 \le \lambda_j \le 1, \quad j = 1, \ldots, M,$$

$$\sum_{j=1}^{M} \lambda_j \pi_k^{(j)} = \pi_k,$$

$$0 \leq \pi_k^{(j)} \leq 1, \quad k \in U, j = 1, \ldots, M,$$
$$\sum_{k \in U} \pi_k^{(j)} = n, \quad j = 1, \ldots, M.$$

We then select one of the set of quantities of $\pi_k^{(j)}, k \in U,$ with probabilities $\lambda_j$ , $j = 1, \ldots, M$. After this first step, any design can be used to select a sample with inclusion probabilities $\pi_k^{(j)}$ or the splitting step can be applied again.

Deville and Tillé (1998) showed that the splitting method defines new sampling designs such as the minimum support design, the splitting into simple random sampling, the pivotal method, and the eliminatory method (Tillé, 2006).

### *2.6. Brewer sampling design*

Brewer (1963) proposed a design for selecting a sample of size $n = 2$. The properties of this design were studied by Rao and Bayless (1969), Rao and Singh (1973), Sadasivan and Sharma (1974), and Cassel et al. (1993). Brewer (1975) generalized this design to any sample size (Brewer and Hanif, 1983, p. 26). This method is a draw by draw procedure, that is, a sample can be selected in $n$ steps. In this section, we show that this design is a particular case of the splitting method.

For the sake of simplicity, only the first step of the method is given. First, consider

$$\lambda_j = \left\{ \sum_{k=1}^{N} \frac{\pi_k(n - \pi_k)}{1 - \pi_k} \right\}^{-1} \frac{\pi_j(n - \pi_j)}{1 - \pi_j}.$$

Secondly, compute

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

The first-order inclusion probabilities are indeed given by $\pi_k$ because

$$\sum_{j=1}^{N} \lambda_j \pi_k^{(j)} = \pi_k.$$

At each step of the method, a unit is selected. Moreover, it is not necessary to compute all the $\pi_k^{(j)}$, as only the selected $\pi_k^{(j)}, k \in U,$ need to be computed.

### *2.7. Maximum entropy or conditional Poisson sampling design*

The maximum entropy design and the conditional Poisson design are the same design obtained from two different perspectives. The maximum entropy design is the design with fixed sample size that maximizes the entropy given in (4) for all the samples of fixed sample size $n$ subject to given inclusion probabilities $\pi_k, k \in U$. Hájek (1981) proposed to implement it by using a Poisson rejective procedure, that is, by reselecting samples by means of a Poisson sampling design until a fixed sample size is obtained. A rejective procedure consists in conditioning Poisson sampling design with respect to

a fixed sample size. Consider a Poisson sampling design with inclusion probabilities $\widetilde{\pi}_k$ and a random sample size $\tilde{n}$. This sampling design can be written as follows:

$$P(s) = \left[\prod_{k\in s} \frac{\widetilde{\pi}_k}{1-\widetilde{\pi}_k}\right]\left[\prod_{k\in U}(1-\widetilde{\pi}_k)\right].$$

The conditional Poisson sampling design is then given by

$$p(s) = P(s|\widetilde{n}_s = n) = \frac{P(s)}{\sum_{s\in\mathcal{S}_n} P(s)}, s \in \mathcal{S}_n,$$

where $n$ is fixed and $\mathcal{S}_n$ is the set of all the samples of size $n$.

Conditional Poisson sampling can be implemented by using a rejective sampling procedure. Samples are selected with Poisson sampling and inclusion probability $\widetilde{\pi}_k$ until a fixed sample size $n$ is obtained. However, more efficient algorithms, such as a draw by draw procedure or a sequential procedure, are described for instance in Tillé (2006, pp. 90–95). The main difficulty is that the inclusion probabilities $\pi_k$ of the design are different from the $\widetilde{\pi}_k$. Hájek (1964) proposed approximations for the inclusion probabilities (see also Brewer and Hanif, 1983, p. 40).

Chen et al. (1994) proposed an algorithm that allows us to derive the inclusion probabilities of the conditional Poisson sampling $\pi_k$ from the inclusion probabilities of the Poisson sampling design $\widetilde{\pi}_k$. In an unpublished manuscript available from the author, Deville (2000) improved this algorithm and derived the following recursive formula:

$$\pi_k(\widetilde{\boldsymbol{\pi}}, n) = n\frac{\widetilde{\pi}_k(1-\widetilde{\pi}_k)^{-1}\left[1-\pi_k(\widetilde{\boldsymbol{\pi}}, n-1)\right]}{\sum_{\ell\in U}\widetilde{\pi}_\ell(1-\widetilde{\pi}_\ell)^{-1}\left[1-\pi_\ell(\widetilde{\boldsymbol{\pi}}, n-1)\right]},$$

where $\widetilde{\boldsymbol{\pi}}$ is the vector of inclusion probabilities $\widetilde{\pi}_k$.

This recursive equation allows us to compute $\pi_k$ from $\widetilde{\pi}_k$ easily. Deville (2000) also proposed that a modified Newton method be used to compute the $\widetilde{\pi}_k$ from the given inclusion probability vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$. The recursive equation is given by

$$\widetilde{\boldsymbol{\pi}}^{(i+1)} = \widetilde{\boldsymbol{\pi}}^{(i)} + \boldsymbol{\pi} - \boldsymbol{\pi}(\widetilde{\boldsymbol{\pi}}, n), \quad \text{for } i = 0, 1, 2, \ldots,$$

where $\widetilde{\boldsymbol{\pi}}^{(0)} = \boldsymbol{\pi}$.

Deville (2000) also proposed a recursive relation for computing the joint inclusion probabilities:

$$\pi_{k\ell}(\widetilde{\boldsymbol{\pi}}, n) = \frac{n(n-1)\exp\lambda_k \exp\lambda_\ell \left[1-\pi_k(\widetilde{\boldsymbol{\pi}}, n-2)-\pi_\ell(\widetilde{\boldsymbol{\pi}}, n-2)+\pi_{k\ell}(\widetilde{\boldsymbol{\pi}}, n-2)\right]}{\sum_{i\in U}\sum_{\substack{j\in U\\ i\neq j}} \exp\lambda_i \exp\lambda_j \left[1-\pi_i(\widetilde{\boldsymbol{\pi}}, n-2)-\pi_j(\boldsymbol{\lambda}, \mathcal{S}_{n-2})+\pi_{ij}(\widetilde{\boldsymbol{\pi}}, n-2)\right]},$$

Additional developments on conditional Poisson sampling are given in Chen et al. (1994), Chen and Liu (1997), Chen (1998, 2000), Deville (2000), Jonasson and Nerman (1996), Aires (1999, 2000), Bondesson et al. (2004), Traat et al. (2004), and Tillé (2006).

### *2.8. Order sampling*

Order sampling designs, developed by (Rosén 1997a, 1997b), are based upon an idea introduced by Ohlsson (1990a). The advantage of order sampling designs is their

simplicity. Let $\pi_k$ be the target first inclusion probability of unit $k$. Consider a positive size variable $x_k > 0$ known for the whole population. The target inclusion probability $\pi_k$ is proportional to $x_k$ and computed as in (1). We generate $N$ uniform random numbers $\omega_k$ in [0,1] and the $n$ units that have the smallest values $\omega_k/\pi_k$ are selected. Other distributions for generating the random numbers can also be used, such as exponential distribution (Hájek, 1964) or Pareto (Rosén 1997a, 1997b) distribution. The main drawback of the method is that the inclusion probabilities are not exactly equal to $\pi_k$. Additional development on order sampling are given in Aires (1999, 2000), Ohlsson (1998), Rosén (2000), Matei and Tillé (2007), and Rosén (1995).

## 3. Point estimation in unequal probability sampling without replacement

We are often interested in estimating population totals of several characteristics of interest. It is therefore possible that some characteristics may not be related to the inclusion probabilities $\pi_k$. In this situation, Rao (1966) recommended the use of the following unweighted estimator

$$\widehat{Y}_{\mathrm{u}} = \frac{N}{n}\sum_{k\in s} y_k. \tag{6}$$

The design bias of this estimator is

$$\mathrm{bias}(\widehat{Y}_{\mathrm{u}}) = \frac{N}{n}\sum_{k\in U} y_k\pi_k - \sum_{k\in U} y_k = \frac{N^2}{n}\frac{1}{M}\sum_{k\in U}(y_k - \widehat{Y})\Big(\pi_k - \frac{n}{N}\Big),$$

which is proportional to the covariance between $y_k$ and $\pi_k$. Thus, this bias is zero when $y_k$ and $\pi_k$ are uncorrelated. Rao (1966) showed that $\widehat{Y}_{\mathrm{u}}$ is on average more accurate than $\widehat{Y}_{\pi}$ because the average variance of $\widehat{Y}_{\mathrm{u}}$ is smaller under the following superpopulation model $\xi$,

$$y_k = \mu + \varepsilon_k, \tag{7}$$

with $\mathrm{E}_\xi(\varepsilon_k|\pi_k) = 0$, $\mathrm{E}_\xi(\varepsilon_k^2|\pi_k) = \sigma^2$, and $\mathrm{E}_\xi(\varepsilon_k\varepsilon_\ell|\pi_k) = 0$, where $\mathrm{E}_\xi(.)$ denotes the expectation under the superpopulation model $\xi$.

Amahia et al. (1989) considered the following linear combination of $\widehat{Y}_{\mathrm{u}}$ and $\widehat{Y}_{\pi}$

$$\widehat{Y}_a = (1-\rho)\widehat{Y}_{\mathrm{u}} + \rho\widehat{Y}_{\pi},$$

where $\rho$ is the observed correlation between $y_k$ and $\pi_k$. This estimator gives more weights to $\widehat{Y}_{\pi}$ when $y_k$ and $\pi_k$ are highly correlated.

The Hájek (1971) estimator, given by

$$\widehat{Y}_{\mathrm{H}} = N\left(\sum_{k\in s}\frac{1}{\pi_k}\right)^{-1}\sum_{k\in s}\frac{y_k}{\pi_k}, \tag{8}$$

is an alternative estimator often used in unequal probability sampling. The estimator $\widehat{Y}_{\mathrm{H}}$ is approximately design unbiased. It should be used when $y_k$ and $\pi_k$ are uncorrelated because its variance may be small when $y_k$ follows model (7) (Särndal et al., 1992,

p. 258). This estimator is often used in practice because it is a weighted average with the sum of weights equal to $N$. This property is particularly useful for the estimation of counts that have to add up to a given constant. Note that with count estimation, the characteristic of interest might not be correlated with $\pi_k$.

When $y_k$ and $\pi_k$ are correlated, $\widehat{Y}_{\mathrm{u}}$ may not be efficient and therefore the $\pi$-estimator should be used instead. When $y_k$ and $\pi_k$ are uncorrelated, $\widehat{Y}_{\mathrm{u}}$ and $\widehat{Y}_{\mathrm{H}}$ should be used. Therefore, the choice of a point estimator should be driven by the correlation between $y_k$ and $\pi_k$ and the $\pi$-estimator should not be blindly used. Basu (1971) gave a famous example, where a circus owner wants to estimate the total weight of his 50 elephants. A sample of size one is selected with inclusion probabilities that are uncorrelated with the weight of each elephant: $\pi_1 = 99/100$ for Dumbo, the average elephant, and $\pi_k = 1/4900$ for the other elephants. Not surprisingly, Dumbo is selected. Let $y_1$ denote its weight. To estimate the total weight, a sensible estimate is $\widehat{Y}_{\mathrm{H}} = \widehat{Y}_{\mathrm{u}} = Ny_1$, which is different from the $\pi$-estimator $\widehat{Y}_{\pi} = y_1 100/99$.

Note that the variance estimator in (3) can be used to derive variance estimators for $\widehat{Y}_{\mathrm{u}}$, $\widehat{Y}_a$, and $\widehat{Y}_{\mathrm{H}}$. By substituting $y_k\pi_k N/n$ for $y_k$ in (3), we obtain a design unbiased estimator for the variance of $\widehat{Y}_{\mathrm{u}}$ when $\pi_{k\ell} > 0$. By substituting $y_k\pi_k/(n/N(1-\rho)+\rho\pi_k)$ for $y_k$ in (3), we obtain an approximately design unbiased estimator for the variance of $\widehat{Y}_a$ when $\pi_{k\ell} > 0$. By substituting $y_k - \widehat{Y}_{\mathrm{H}}$ for $y_k$ in (3), we obtain a approximately design unbiased estimator for the variance of $\widehat{Y}_{\mathrm{H}}$ when $\pi_{k\ell} > 0$.

The choice of the size variable should be driven by the correlation between the variable of interest and the size variable. Ideally, the size variable should be highly correlated with the variable of interest. However, in practice, we have severale variables of interest and the size variable might be not correlated with all the variables of interest. In this situation, we recommend to use the simple mean (6) or the Hájek estimator (8) to estimate a total.

## 4. Variance estimators free of joint inclusion probabilities

Exact joint inclusion probabilities may be difficult or impossible to calculate. Futhermore, the double sum in (3) makes the Sen–Yates–Grundy estimator computationally intensive when the sample size is large. It is also inconceivable to provide these probabilities in released data sets, as the set of joint inclusion probabilities is a series of $n(n-1)/2$ values. Suppose that the sampling design uses single-stage stratified sampling with unequal probabilities within each stratum. Let $U_1, \ldots, U_H$ denote the strata. Suppose that a sample $s_h$ of size $n_h$ is selected without replacement within each stratum $U_h$ of size $N_h$. In this situation, we can estimate the variance of $\widehat{Y}_{\pi}$ approximately by

$$\widehat{\mathrm{var}}^{*}\left(\widehat{Y}_{\pi}\right) = \sum_{k\in s} \alpha_k \widehat{e}_k^2, \tag{9}$$

which is free of the $\pi_{l\ell}$. The $\widehat{e}_k$ are the residuals of weighted least squares given by

$$\widehat{e}_k = \frac{y_k}{\pi_k} - \sum_{h=1}^{H} \widehat{B}_h z_{kh},$$

and $\widehat{B}_h$ is the weighted least squares regression coefficient given by

$$\widehat{B}_h = \left(\sum_{k\in s} \lambda_k z_{kh}^2\right)^{-1} \sum_{k\in s} \lambda_k z_{kh} \frac{y_k}{\pi_k},$$

where $z_{kh} = 1$ if $k \in U_h$ and otherwise $z_{kh} = 0$. The choice of $\alpha_k$ and $\lambda_k$ depends on the value of $n_h$ and on the sampling design implemented. Several choices are possible for the constants $\alpha_k$ and $\lambda_k$. A simple choice is $\alpha_k = \lambda_k = 1$, which gives the naive variance estimator under sampling with replacement given in (5). However, this approach usually leads to overestimation of the variance for large sampling fraction. When $\alpha_k = 1 - \pi_k(n_h - 1)/n_h$ for $k \in U_h$ and $\lambda_k = 1$, (9) reduces to the Hartley and Rao (1962) variance estimator. When $\alpha_k = \lambda_k = (1 - \pi_k)n_h/(n_h - 1)$, for $k \in U_h$, (9) reduces to the Hájek (1964) variance estimator.

For the randomized systematic sampling method, Hartley and Rao (1962) showed that var $(\widehat{Y}_\pi)$ reduces to

$$\text{var}\,(\widehat{Y}_\pi) \approx \sum_{h=1}^{H} \sum_{k\in U_h} \pi_k \left(1 - \frac{n_h - 1}{n_h}\pi_k\right)\left(\frac{y_k}{\pi_k} - \frac{Y}{n}\right)^2 \tag{10}$$

for fairly large $N_h$ and for small sampling fractions. Therefore, (9) will be a consistent estimator of (10) under the randomized systematic design, when $\alpha_k = 1-\pi_k(n_h-1)/n_h$ for $k \in U_h$, $\lambda_k = 1$. This choice is recommended when $n_h$ is small and $N_h$ is large, or when $n_h$ is large and $n_h/N_h$ is negligible.

Assuming $d_h = \sum_{\ell\in U_h} \pi_\ell(1 - \pi_\ell) \to \infty$, Hájek (1964) derived an approximation to $\pi_{k\ell}$ under maximum entropy sampling. By substituting this expression into (3), we have

$$\text{var}\,(\widehat{Y}_\pi) = \sum_{k\in U} \pi_k(1 - \pi_k)e_k^2,$$

with

$$e_k = \frac{y_k}{\pi_k} - \sum_{h=1}^{H} B_h z_{kh},$$

where $B_h$ is the following population weighted least squares regression estimate

$$B_h = \left(\sum_{k\in U}(1 - \pi_k)z_{kh}^2\pi_k\right)^{-1} \sum_{\ell\in U}(1 - \pi_\ell)z_{\ell h}y_\ell\pi_\ell.$$

Therefore, (9) will be a consistent estimator of (10) under maximum entropy sampling, when $\alpha_k = \lambda_k = 1 - \pi_k$ and $d_h \to \infty$. This choice is recommended when $n_h$ is large and the sampling fraction is not small. Berger (2007) showed that this choice gives a consistent estimator for the variance under the Rao–Sampford sampling design when $d_h \to \infty$, $H$ bounded, and none of the $\pi_k$ less than 1 approach 1 asymptotically. Berger (2005a) showed that this choice is suitable for the Chao (1982) sampling design.

Other choices for $\alpha_k$ and $\lambda_k$ have been proposed in literature. When $\alpha_k = \lambda_k = (1 - \pi_k)\log(1 - \pi_k)/\pi_k$, (9) reduces to the Rosén (1991) estimator. When $\alpha_k = (1 - \pi_k)n_h(n_h - 1)\sum_{k\in s_h}(1 - \pi_k)\left(\sum_{k\in U_k}\pi_k(1 - \pi_k)\right)^{-1}$, (9) gives the Berger (1998)

estimator. If $\alpha_k = \lambda_k = (1-\pi_k)^{-1}\left[1 - d_h^{-2}\sum_{\ell\in s_h}(1-\pi_\ell)\right]$ for $k \in U_h$, (9) gives the Deville (1999) variance estimator. Brewer (2002, Chapter 9) proposed two alternative choices for $\alpha_k$ and $\lambda_k$. Simulation studies by Brewer (2002), Haziza et al. (2004), Matei and Tillé (2005), and Henderson (2006) showed that (9) is an accurate estimator for various choices of $\alpha_k$ and $\lambda_k$. The variance estimator (9) may have a smaller mean square error than the exactly unbiased Sen–Yates–Grundy estimator in (3).

Berger (2005a) showed that (9) can be easily computed when $\alpha_k = \lambda_k$, as (9) reduces to $\widehat{\text{var}}^*(\widehat{Y}_\pi) = n\widehat{\sigma}_\varepsilon^2$, where $\widehat{\sigma}_\varepsilon^2$ is the observed residual variance of the regression

$$y_k^* = \sum_{h=1}^{H} \beta_h z_{\ell h}^* + \varepsilon_k$$

fitted with ordinary least squares, where the $\varepsilon_k$ are independent normal random variables with mean 0 and variances $\sigma_\varepsilon^2$, $y_k^* = y_k \pi_k^{-1}\alpha_k^{1/2}$ and $z_k^* = z_k \pi_k^{-1}\alpha_k^{1/2}$.

## 5. Variance estimation of a function of means

Assume that the parameter of interest $\theta$ can be expressed as a function of means of $Q$ survey variables, that is, $\theta = g(\mu_1, \ldots, \mu_Q)$, where $g(.)$ is a smooth differentiable function (Shao and Tu, 1995, Chapter 2), and $\mu_q$ is the finite population mean of the $q$th survey variables. This definition includes parameters of interest arising in common survey applications such as ratios, subpopulation means, and correlation and regression coefficients. It excludes parameters such as L-statistics (Shao, 1994) and coefficients of logistic regression, which cannot be expressed as function of means. The parameter $\widehat{\theta}$ can be estimated by the substitution estimator $\widehat{\theta} = g(\widehat{\mu}_{1H}, \ldots, \widehat{\mu}_{QH})$, in which $\widehat{\mu}_{qH}$ is the Hájek (1971) estimator of a the $q$th mean.

The variance of $\widehat{\theta}$ can be estimated by the linearized variance estimator (Rabinson and Särndal, 1983) given by

$$\widehat{\text{var}}(\widehat{\theta})_{\text{L}} = \nabla(\widehat{\boldsymbol{\mu}})'\widehat{\Sigma}\nabla(\widehat{\boldsymbol{\mu}}),$$

where

$$\widehat{\Sigma} = \frac{1}{N^2}\sum_{k\in s}\sum_{\ell\in s}\left(\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_k\pi_\ell\pi_{k\ell}}\right)(\mathbf{y}_k - \widehat{\boldsymbol{\mu}})\,(\mathbf{y}_\ell - \widehat{\boldsymbol{\mu}})',$$

$$\nabla(\mathbf{x}) = \left(\frac{\partial g(\boldsymbol{\mu})}{\partial\mu_1}, \ldots, \frac{\partial g(\boldsymbol{\mu})}{\partial\mu_Q}\right)'_{\mu=\mathbf{x}},$$

$\mathbf{y}_k = (y_{1k}, \ldots, y_{Qk})'$, $\nabla(\mathbf{x})$ denotes the gradient of $g(\cdot)$ at $\mathbf{x} \in \mathbb{R}^Q$, $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_{1H}, \ldots, \widehat{\mu}_{QH})'$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_Q)'$.

Customary jackknife variance estimators (Shao and Tu, 1995; Wolter, 1985) are not always consistent under unequal probability sampling without replacement (Demnati and Rao, 2004). Campbell (1980) proposed a generalized jackknife variance estimator that allows us to estimate the variance for unequal probability sampling and stratification. Campbell's generalized jackknife is given by

$$\widehat{\text{var}}\,(\widehat{\theta}) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}u_k u_\ell,$$

where

$$u_j = (1 - w_j)(\widehat{\theta} - \widehat{\theta}_{(j)}),$$

$$w_j = \pi_j^{-1} \left( \sum_{k \in s} \pi_k^{-1} \right)^{-1},$$

$$\widehat{\theta}_{(j)} = g(\widehat{\mu}_{1H(j)}, \ldots, \widehat{\mu}_{QH(j)}),$$

$$\widehat{\mu}_{qH(j)} = N \left( \sum_{k \in s} \frac{\delta_{kj}}{\pi_k} \right)^{-1} \sum_{k \in s} \frac{\delta_{kj} y_k}{\pi_k},$$

and $\delta_{kj} = 1$ if $k = j$ and $\delta_{kj} = 0$ otherwise. Berger and Skinner (2005) gave regularity conditions under which the generalized jackknife is consistent. They also showed that the generalized jackknife may be more accurate than the customary jackknife estimators. Berger (2007) proposed an alternative consistent jackknife estimator that is free of joint inclusion probabilities.

Many surveys use single imputation to handle item nonresponse. Treating the imputed values as if they were true values and then estimating the variance using standard methods may lead to serious underestimation of the variance when the proportion of missing values is large (Rao and Shao, 1992; Särndal, 1992). One can use the Rao–Shao method which consists of adjusting the imputed values whenever a responding unit is deleted. Berger and Rao (2006) showed that this method gives a consistent generalized jackknife variance estimator under uniform response.

## 6. Balanced sampling

### *6.1. Definition*

A design is balanced if the $\pi$-estimators for a set of auxiliary variables are equal to the known population totals of auxiliary variables. Balanced sampling can be viewed as a calibration method embedded into the sampling design. Yates (1949) advocated the idea of respecting the means of known variables in probability samples. Yates (1946) and Neyman (1934) described methods of balanced sampling limited to one variable and to equal inclusion probabilities. The use of balanced sampling was recommend by Royall and Herson (1973) for protecting inference against misspecified models. More recently, several partial solutions were proposed by Deville et al. (1988), Deville (1992), Ardilly (1991), and Hedayat and Majumdar (1995). Valliant et al. (2000) surveyed some existing methods.

The cube method (Deville and Tillé, 2004) is a general method of balanced sampling with equal or unequal inclusion probabilities. Properties and application of this method were studied in Deville and Tillé (2004), Chauvet and Tillé (2006), Tillé and Favre (2004, 2005), Berger et al. (2003), and Nedyalkova and Tillé (2008). The cube method was used to select the rotation groups of the new French census (Bertrand et al., 2004; Dumais and Isnard, 2000; Durr and Dumais, 2002) and the selection of the French master sample (Christine, 2006; Christine and Wilms, 2003; Wilms, 2000). Deville and Tillé (2005) proposed a variance estimator for balanced sampling. Deville (2006) also proposed to use balanced sampling for the imputation of item nonresponse. The cube

AQ:2 method can be implemented in practice by SAS or R procedures (Chauvet and Tillé, 2005; Rousseau and Tardieu, 2004; Tardieu, 2001; Tillé and Matei, 2007).

Balancing is used when auxiliary information is available at the design stage. When balanced sampling is used, the Horvitz–Thompson weights are also calibration weights. Calibration after sampling is therefore not necessary. Balancing also provide more stable estimators as these weights do not depend on the sample.

### *6.2. Balanced sampling and the cube method*

Suppose that the values of $p$ auxiliary variables $x_1, \ldots x_p$ are known for every unit of the population. Let $\mathbf{x}_k = (x_{k1} \cdots x_{kj} \cdots x_{kp})'$ be the vector of the $p$ auxiliary variables on unit $k$. For a set of given inclusion probabilities $\pi_k$, a design $p(.)$ is said to be balanced with respect to the auxiliary variables $x_1, \ldots, x_p$, if and only if it satisfies the balancing equations given by

$$\sum_{k \in s} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k. \tag{11}$$

Balanced sampling generalizes several well-known methods. For instance, if $\mathbf{x}_k = \pi_k$, then (11) is a fixed size constraint. It can be shown that, if the auxiliary variables are the indicator variables of strata, a stratified sampling design is balanced on these indicator variables.

However, it is often not possible to find a sample such that (11) holds, for example, when the right-hand side of (11) is an integer. Hence, an exactly balanced design often does not exist. For example, if $x_1 = 1$, $x_2 = 1$, $x_3 = 1$, $x_1 = 5$, and $\pi_k = 1/2$, for $i = 1, 2, 3, 4$, the balancing equation becomes

$$\sum_{k \in s} 2x_k = \sum_{k \in U} x_k = 11, \tag{12}$$

which cannot hold. The aim is thus to select an exact balanced sample if possible, and an approximately balanced sample otherwise.

The name "cube method" comes from the geometrical representation of a sampling design. A sample can be written as a vector $\mathbf{s} = (s_1, \ldots, s_N) \in \mathbb{R}^N$ of indicator variables $s_k$ that take the value 1 if the unit is selected and 0 otherwise. Geometrically, each vector $\mathbf{s}$ can be viewed as one of the $2^N$ vertices of a $N$-cube in $\mathbb{R}^N$. A design consists thus in allocating a probability $p(.)$ to each vertex of the $N$ cube in such a way that the expectation of $\boldsymbol{s}$ is equal to the inclusion probability vector $\boldsymbol{\pi}$, that is,

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in \mathcal{S}} p(s)\mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} \in \mathbb{R}^N$ is the vector of inclusion probabilities. Thus, selecting a sample consists in choosing a vertex (a sample) of the $N$-cube that is balanced.

The balancing equations in (11) can also be written as

$$\sum_{k \in U} \mathbf{a}_k s_k = \sum_{k \in U} \mathbf{a}_k \pi_k \text{ with } s_k \in \{0, 1\}, k \in U,$$

where $\mathbf{a}_k = \mathbf{x}_k/\pi_k, k \in U$. The balancing equations define an affine subspace in $\mathbb{R}^N$ of dimension $N - p$ denoted $Q$. The subspace $Q$ can be written as $\boldsymbol{\pi} + \text{Ker}\mathbf{A}$, where $\text{Ker}\mathbf{A} = \{\boldsymbol{u} \in \mathbb{R} | \mathbf{A}\boldsymbol{u}\}$ and $\mathbf{A} = (\mathbf{a}_1 \cdots \mathbf{a}_n \cdots \mathbf{a}_N)$.

It is possible to geometrically represent the situation when (12) does not hold. When the vertices of the intersection between the cube and $Q$ are also vertices of the cube, as in Fig. 1, a balanced sample can be selected. When the vertices of the intersection between the cube and $Q$ are not vertices of the cube, as in Fig. 2, it is not possible to select an exact balanced sample. In this situation, only an approximately balanced sample can be selected (see Section 6.4).

### *6.3. The flight phase*

The cube method is made up of two parts: the flight phase and the landing phase. *The flight phase*, described in Algorithm 1 below, is a random walk which begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace $Q$. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. There are several ways to implement this algorithm. Chauvet and AQ:3

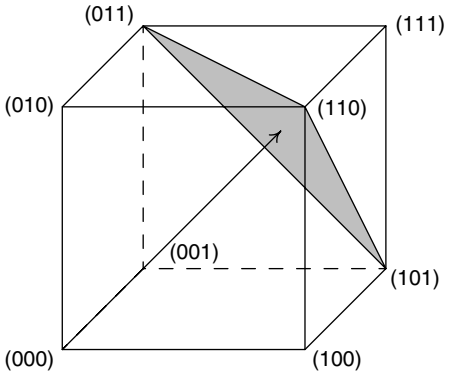


Fig. 1. Fixed size constraint of size 2: an exact balanced sample always exists.

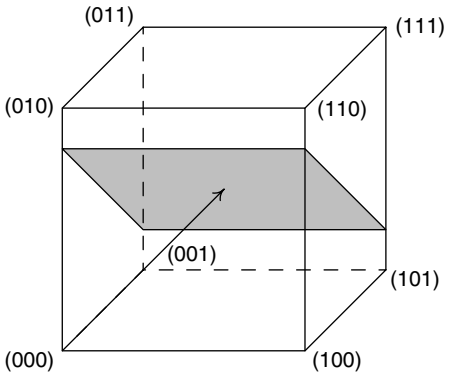


Fig. 2. The balanced constraints are such that an exact balanced sample does not exist.

Tillé (2006) proposed a fast algorithm whereby the calculation time increases linearly with the population size.

**Algorithm 1:** Flight phase of the cube method

---

First initialize with $\boldsymbol{\pi}(0) = \boldsymbol{\pi}$.
Next, at time $t = 1, \ldots, T$,

1. Generate any vector $\boldsymbol{u}(t) = [u_k(t)] \neq 0$ so that
   *(i)* $\boldsymbol{u}(t)$ is in the kernel of matrix $\boldsymbol{A}$
   *(ii)* $u_k(t) = 0$ if $\pi_k(t)$ is an integer.
2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values so that
   $0 \leq \boldsymbol{\pi}(t) + \lambda_1(t)\boldsymbol{u}(t) \leq 1$,
   $0 \leq \boldsymbol{\pi}(t) - \lambda_2(t)\boldsymbol{u}(t) \leq 1$.
3. Compute $\boldsymbol{\pi}(t+1) = \begin{cases} \boldsymbol{\pi}(t) + \lambda_1^*(t)\boldsymbol{u}(t) & \text{with a proba } q_1(t) \\ \boldsymbol{\pi}(t) - \lambda_2^*(t)\boldsymbol{u}(t) & \text{with a proba } q_2(t), \end{cases}$
   where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)\}$.

---

### *6.4. Landing phase*

*The landing phase* begins at the end of the flight phase. If a sample is not obtained at the end of the flight phase, a sample is selected as close as possible to the constraint subspace. At the end of the flight phase, Algorithm 1 stops on a vertex denoted $\boldsymbol{\pi}^*$ of the intersection between the cube and $Q$. It is possible to show that

$$\text{card } U^* = \text{card } \left\{k \in U | 0 < \pi_k^* < 1\right\} = q \leq p,$$

which means that the number of noninteger elements of $\boldsymbol{\pi}^*$ is smaller or equal to the number of balancing variables. The aim of the landing phase is to find a random sample $\boldsymbol{s}$ so that $\text{E}(\boldsymbol{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$ and which is almost balanced.

Two solutions can be used to select the sample. The first solution consists of enumerating all the samples that are consistent with $\boldsymbol{\pi}^*$, a sample $\mathbf{s}$ being consistent if $s_k = \pi_k^*$ when $\pi_k^*$ is an integer. Then, a cost $C(\mathbf{s})$ is attached at each sample. This cost is equal to zero when the sample is balanced and which increases when the sample moves away from the subspace $Q$. Deville and Tillé (2004) proposed several $C(\mathbf{s})$. By a method of linear programming, it is possible to find a sampling design on the consistent samples that satisfies the inclusion probability $\boldsymbol{\pi}^*$ and which minimizes the average cost. Finally, a sample is selected at random, following this sampling design. This method can be used with a number of balancing variables that are less than 15 because it is necessary to enumerate the $2^{15}$ samples.

The second method can be used when the number of auxiliary variables is too large for the solution to be obtained by a simplex algorithm. At the end of the flight phase, an auxiliary variable can be dropped out. Next, one can return to the flight phase until it is no longer possible to "move" within the constraint subspace. The constraints are thus successively relaxed until the sample is selected.

**Author Queries**

**AQ:1** Please confirm whether the closing parenthesis in [1,0) in the sentence “A simple method to seleet a sample...” changed to square bracket is ok.‘

**AQ:2** Please define “SAS” in the sentence “Deville (2006) also proposed to use balanced...”

**AQ:3** Please confirm “Algorithm 2 changed to Algorithm 1” is ok.