Electronic Journal of Statistics
Vol. 6 (2012) 1477–1489
ISSN: 1935-7524
DOI: 10.1214/12-EJS719

Size constrained unequal probability sampling with a non-integer sum of inclusion probabilities

Anton Grafström

Swedish University of Agricultural Sciences, Umeå, Sweden
e-mail: Anton.Grafstrom@slu.se

Lionel Qualité, Yves Tillé and Alina Matei

University of Neuchâtel, Switzerland
e-mail: Lionel.Qualite@unine.ch; Yves.Tille@unine.ch; Alina.Matei@unine.ch

Abstract: More than 50 methods have been developed to draw unequal probability samples with fixed sample size. All these methods require the sum of the inclusion probabilities to be an integer number. There are cases, however, where the sum of desired inclusion probabilities is not an integer. Then, classical algorithms for drawing samples cannot be directly applied. We present two methods to overcome the problem of sample selection with unequal inclusion probabilities when their sum is not an integer and the sample size cannot be fixed. The first one consists in splitting the inclusion probability vector. The second method is based on extending the population with a phantom unit. For both methods the sample size is almost fixed, and equal to the integer part of the sum of the inclusion probabilities or this integer plus one.

AMS 2000 subject classifications: Primary 62D05.
Keywords and phrases: Survey sampling, maximum entropy, splitting method.

Received February 2012.

1. Introduction

Unequal probability sampling with fixed sample size is an intricate problem. At least 50 methods are described in Brewer and Hanif (1983) and Tillé (2006) to draw an unequal probability sample with fixed sample size. All these methods assume the sum of the inclusion probabilities to be an integer number. There are cases, however, where this sum is not an integer. Two main examples where the sum of inclusion probabilities is not an integer are given below. A first example is that of sampling with probabilities proportional to size from a population divided into domains. Inclusion probabilities within domains often do not sum to integer numbers. Consequently, when one wants to control the sample size within domains, one usually uses rounding algorithms to obtain integer sample sizes for all domains, maintain the requested total sample size, and then use a stratified sampling algorithm to select the sample. However, this can become

problematic when there is a large number of domains with small expected sample sizes, as is commonly the case in business surveys. In this kind of survey, the population is usually divided into size classes and activity sectors. Results are published by economic sectors of activity, according to a classification defined within each country. Size classes, in terms of revenue or of number of employees, is a secondary domain of interest for publications. More importantly, business sizes have to be taken into account in order to build an efficient sampling design. Sample sizes are then controlled for cells that are the intersection of these attributes. This is done in order to ensure that fixed sample sizes are respected both at the size class level and at the activity sector level, and also that further aggregations will not be hindered by accidentally empty sample cells. The proportionality relation between inclusion probabilities and business sizes is degraded by the large number of roundings and the resulting relative deviation from the original inclusion probabilities, that can be important for small domains. Another example is that of bootstrap procedures. Antal and Tillé (2011) have proposed a bootstrap method where units are re-sampled from the initial sample using the original inclusion probabilities. However, the sum of these probabilities within the bootstrap sample is usually not integer.

The case where the sum of inclusion probabilities is not an integer number was recently discussed by Bondesson and Grafström (2011). They proposed a generalization of the Sampford (1967) method to the case where the sum of the inclusion probabilities is not an integer. In their solution, the selection of one unit of the population is dealt with in a special way, and the final sample size is equal to the integer directly below the sum of inclusion probabilities, or to the integer directly above it. In this paper, we describe general solutions to overcome the problem when the sum of the inclusion probabilities is not an integer. All fixed size sampling designs can be, through these solutions, generalized to inclusion probabilities that do not sum to an integer. We give practical procedures to do so, and in particular to implement a maximum entropy design (see Hájek, 1981).

The paper is organized as follows. In Section 2 we present the first method based on splitting the inclusion probability vector into two new inclusion probability vectors. For this method we present two different algorithms for the splitting. One is based on the πps procedure for calculating inclusion probabilities and the other one allows sampling with maximum entropy. Differences between these two splitting algorithms are illustrated with a small example. The second method, based on an augmented population, is presented in Section 3. Estimation, with a small example, is shortly treated in Section 4. We comment on some applications in Section 5. Finally, in Section 6 we discuss the interest of the different methods.

2. First general solution by splitting the inclusion probability vector

2.1. Splitting into two fixed size designs

Assign a number $\pi_k \in [0, 1]$ to all units k of a finite population U, and suppose that one wants to randomly select a subset s of U with inclusion probabilities

contained in the vector $\boldsymbol{\pi} = (\pi_k)_{k\in U}$. The value $\eta = \sum_{k\in U} \pi_k$ gives the expected size of the selected sample. When η is an integer, there exist many methods (see Brewer and Hanif, 1983; Tillé, 2006) of selection whereby only samples of size η can be selected.

When η is not an integer, we may want to use a method that enables us to select a sample with a size that is close to η while respecting the inclusion probabilities π_k. More precisely, the size of the selected sample should be either equal to n or $n+1$, where n is the integer such that $n \leq \eta < n+1$. We are looking for implementations of probability distributions P on the subsets s of U such that

$$P(s) > 0 \Rightarrow |s| \in \{n, n+1\},\ s \subset U, \tag{2.1}$$

and

$$\sum_{s\subset U} I_k(s)P(s) = \pi_k,\ k \in U, \tag{2.2}$$

where $|s|$ is the cardinal of s, and $I_k(s) = 1$ if $k \in s$ and 0 otherwise is the sample membership indicator function. These constraints imply that $P(\{|s| = n+1\}) = \eta - n$.

A first possible solution is to describe all probability distributions that satisfy conditions (2.1) and (2.2) through the splitting method developed by Deville and Tillé (1998). These distributions are obtained by constructing two vectors denoted by $\boldsymbol{\pi}^- = (\pi_k^-)_{k\in U}$ and $\boldsymbol{\pi}^+ = (\pi_k^+)_{k\in U}$, such that

$$0 \leq \pi_k^- \leq 1,\ 0 \leq \pi_k^+ \leq 1,\ k \in U, \tag{2.3}$$

$$\pi_k = (1-q)\pi_k^- + q\pi_k^+,\ k \in U, \tag{2.4}$$

$$\sum_{k\in U} \pi_k^- = n,\ \text{and}\ \sum_{k\in U} \pi_k^+ = n+1, \tag{2.5}$$

where $q = \eta - n$. Once these vectors are computed, a realization r of a Bernoulli variable with parameter q is generated. If $r = 1$, a sample s of size $n+1$ is drawn from U using any fixed-size method with inclusion probabilities $(\pi_k^+)_{k\in U}$. Similarly, if $r = 0$, a sample s of size n is drawn from U using any fixed-size method with inclusion probabilities $(\pi_k^-)_{k\in U}$. It is easy to see that this procedure gives a solution to our problem, and that all probability distributions that are solutions can be found through this procedure. We give in Subsections 2.2 and 2.3 two different methods to compute vectors $\boldsymbol{\pi}^+$ and $\boldsymbol{\pi}^-$. The first one mimics the usual probability proportional to size computation of inclusion probabilities (πps procedure); the second one allows implementation of the maximum entropy sampling design.

2.2. Computation based on the πps procedure

Suppose that η is not integer and that n is the integer such that $n < \eta < n+1$. Vectors $\boldsymbol{\pi}^+ = (n+1)\boldsymbol{\pi}/\eta$ and $\boldsymbol{\pi}^- = n\boldsymbol{\pi}/\eta$ satisfy Conditions (2.4) and (2.5) but $\boldsymbol{\pi}^+$ does not necessarily satisfy the second relationship in (2.3).

However, suitable vectors $(\pi_k^-)_{k\in U}$ and $(\pi_k^+)_{k\in U}$ can be found through the usual procedure for probability proportional to size sampling (see Särndal et al., 1992, p. 89). Here the size measure is the original inclusion probability vector $\boldsymbol{\pi}$. For some units with large inclusion probabilities π_k, the quantities $(n+1)\pi_k/\eta$ may be larger than one. The standard procedure is then to assign to these units an inclusion probability equal to one, and to compute proportional to size inclusion probabilities for the remaining units, repeating the operation several times if necessary. The solution can also be found directly in a few steps given in Algorithm 2.1.

Algorithm 2.1 (Direct computation of a πps-inclusion probability vector)**.** *Here, the size measure is the value of π_k, the desired sample size is $n+1$ and the obtained inclusion probabilities are π_k^+, $k\in U$. One gives the general solution to the computation of a πps inclusion probability vector substituting $n+1$ with a suitable size.*

1. *Order the population units so that $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_N$,*
2. *Compute*
$$u_i = (n+2-i)\frac{\pi_i}{\sum_{k=i}^N \pi_k},\ i=1,\ldots,N.$$
3. *Define $A = \{i$ such that $u_\ell \geq 1$ for all $1\leq \ell \leq i\}$.*
4. *Define $\pi_k^+ = 1$ if $k\in A$ and,*
$$\pi_k^+ = (n+1-|A|)\frac{\pi_k}{\sum_{i\notin A}\pi_i} \text{ otherwise, if } |A| < n+1.$$

If $|A| = n+1$ then $\pi_k^+ = 0$ for all $k\notin A$.

We then define $\pi_k^- = (\pi_k - q\pi_k^+)/(1-q)$, for all $k\in U$. As stated in Proposition 2.2, vectors π_k^+ and π_k^- enjoy the required properties and can be used to implement a size constrained sampling design with average size η.

Proposition 2.2. *Vectors $\boldsymbol{\pi}^+$ and $\boldsymbol{\pi}^-$ computed through Algorithm 2.1 satisfy Conditions (2.3), (2.4) and (2.5). Furthermore, we have that*
$$0\leq \pi_k^- \leq \pi_k \leq \pi_k^+ \leq 1, \text{ for all } k\in U.$$

A proof of Proposition 2.2 is given in Appendix.

2.3. Computation for maximum entropy sampling design

Consider that each sample s of U is represented by a vector $\mathbf{s}$ in $\mathbb{R}^N$, $N=|U|$, whose components are equal to the sample membership indicators, so that $\mathbf{s}_k = I_k(s)$, for all $k\in U$. Let $\mathcal{S}$ denote the set of all possible samples of any size $1\leq n\leq N$ of U, and $\mathcal{Q}$ be a subset of $\mathcal{S}$. A maximum entropy sampling design $p(.)$ on $\mathcal{Q}$ is a probability distribution whose support is equal to $\mathcal{Q}$ and that maximizes the entropy function, given by
$$H(p) = -\sum_{\mathbf{s}\in\mathcal{Q}} p(\mathbf{s})\log p(\mathbf{s}).$$

Some of the most frequently used sampling designs, such as simple random sampling, stratified sampling and Poisson sampling, are maximum entropy designs. The use of maximum entropy or high entropy sampling designs is commonly deemed to be desirable on the grounds that these designs retain a high level of randomness, even when they are subject to size and inclusion probabilities constraints. Properties of high entropy sampling designs are discussed amongst others in Hájek (1964); Berger (1998); Brewer and Donadio (2003) and Qualité (2008). If the sampling design is constrained to a vector of given inclusion probabilities $\boldsymbol{\pi}$, then (see Darroch and Ratcliff, 1972)

$$p(\mathbf{s}, \mathcal{Q}, \boldsymbol{\lambda}) = \frac{\exp(\boldsymbol{\lambda}'\mathbf{s}) I(\mathbf{s} \in \mathcal{Q})}{\alpha(\mathcal{Q}, \boldsymbol{\lambda})},$$

where $\alpha(\mathcal{Q}, \boldsymbol{\lambda}) = \sum_{\mathbf{s}\in\mathcal{Q}} \exp(\boldsymbol{\lambda}'\mathbf{s})$, $I(\mathbf{s} \in \mathcal{Q})$ is equal to 1 if $\mathbf{s} \in \mathcal{Q}$ and to 0 otherwise, and the vector-parameter $\boldsymbol{\lambda} \in \mathbb{R}^N$ is such that the inclusion probabilities π_k, $k \in U$ are obtained, i.e.

$$\sum_{\mathbf{s}\in\mathcal{Q}} \mathbf{s} \cdot p(\mathbf{s}, \mathcal{Q}, \boldsymbol{\lambda}) = \boldsymbol{\pi}. \tag{2.6}$$

Such a vector always exists when $\boldsymbol{\pi}$ lies in the interior of the convex hull of $\mathcal{Q}$ (see Brown, 1986, p.74). Degenerate cases where $\boldsymbol{\pi}$ is on the boundary of this set can be accounted for by cautiously allowing some coordinates λ_k to take infinite values. A fast algorithm that allows to compute $\boldsymbol{\pi}$ from $\boldsymbol{\lambda}$ and reciprocally is described amongst others in Tillé (2006, pp. 82-83) in the case of the maximum entropy sampling designs with fixed sample size, also called conditional Poisson sampling.

In the present case, the support $\mathcal{Q}$ is the set of all samples of size n, denoted by $\mathcal{S}_n$ added to the set of all samples of size $n+1$, denoted by $\mathcal{S}_{n+1}$. Hence we have that

$$\alpha(\mathcal{Q}, \boldsymbol{\lambda}) = \alpha(\mathcal{S}_n, \boldsymbol{\lambda}) + \alpha(\mathcal{S}_{n+1}, \boldsymbol{\lambda}).$$

We also have that

$$q = \eta - n = p\left(\{|\mathbf{s}| = n+1\}\right) = \sum_{\mathbf{s}\in\mathcal{S}_{n+1}} p(\mathbf{s}, \mathcal{Q}, \boldsymbol{\lambda}) = \frac{\alpha(\mathcal{S}_{n+1}, \boldsymbol{\lambda})}{\alpha(\mathcal{Q}, \boldsymbol{\lambda})}. \tag{2.7}$$

We can thus write the natural decomposition:

$$\begin{aligned} p(\mathbf{s}, \mathcal{Q}, \boldsymbol{\lambda}) &= \frac{\exp(\boldsymbol{\lambda}'\mathbf{s})}{\alpha(\mathcal{Q}, \boldsymbol{\lambda})} I(\mathbf{s} \in \mathcal{Q}) \\ &= (1-q)\frac{\exp(\boldsymbol{\lambda}'\mathbf{s})}{\alpha(\mathcal{S}_n, \boldsymbol{\lambda})} I(\mathbf{s} \in \mathcal{S}_n) + q\frac{\exp(\boldsymbol{\lambda}'\mathbf{s})}{\alpha(\mathcal{S}_{n+1}, \boldsymbol{\lambda})} I(\mathbf{s} \in \mathcal{S}_{n+1}) \\ &= (1-q)\ p(\mathbf{s}, \mathcal{S}_n, \boldsymbol{\lambda}) + q\ p(\mathbf{s}, \mathcal{S}_{n+1}, \boldsymbol{\lambda}). \end{aligned}$$

It follows that maximum entropy sampling with given inclusion probabilities $\boldsymbol{\pi}$ and support $\mathcal{Q}$ is a mixture of maximum entropy sampling on support $\mathcal{S}_n$ and of maximum entropy sampling on support $\mathcal{S}_{n+1}$ with the same vector-parameter $\boldsymbol{\lambda}$.

Usual numerical approximation procedures to obtain a suitable vector $\boldsymbol{\lambda}$ from $\boldsymbol{\pi}$ (see for example Chen et al., 1994; Aires, 2000; Deville, 2000; Tillé, 2006, p.83) can easily be adapted to the case where $\mathcal{Q} = \mathcal{S}_n \cup \mathcal{S}_{n+1}$. Indeed, these procedures are based on a simplified Newton algorithm to find a solution $\boldsymbol{\lambda}$ of

$$\boldsymbol{\pi} - \boldsymbol{\pi}(n, \boldsymbol{\lambda}) = \mathbf{0},$$

where $\boldsymbol{\pi}(n, \boldsymbol{\lambda}) = \sum_{\mathbf{s} \in \mathcal{S}_n} \mathbf{s} \cdot p(\mathbf{s}, \mathcal{S}_n, \boldsymbol{\lambda})$, and the fact that we can explicitly compute conditional inclusion probabilities $\boldsymbol{\pi}(n, \boldsymbol{\lambda})$ for any integer n and parameter $\boldsymbol{\lambda}$. In order to derive a suitable $\boldsymbol{\lambda}$ from $\boldsymbol{\pi}$, we can use the same idea to find a solution of

$$\boldsymbol{\pi} - q\boldsymbol{\pi}(n+1, \boldsymbol{\lambda}) - (1-q)\boldsymbol{\pi}(n, \boldsymbol{\lambda}) = \mathbf{0}. \tag{2.8}$$

Once $\boldsymbol{\lambda}$ is obtained, we compute inclusion probability vectors $\boldsymbol{\pi}^- = \boldsymbol{\pi}(n, \boldsymbol{\lambda})$ and $\boldsymbol{\pi}^+ = \boldsymbol{\pi}(n+1, \boldsymbol{\lambda})$. These vectors automatically satisfy Conditions (2.3), (2.4) and (2.5). Moreover, with this method, inequalities $\pi_k^- \leq \pi_k \leq \pi_k^+$ hold for all k in U as the inclusion probabilities of conditional Poisson sampling with a given parameter $\boldsymbol{\lambda}$ increase with the size of the samples (see for example Hájek, 1981). Any fixed-size design can then be used with these vectors, or, having already computed $\boldsymbol{\lambda}$, we can very rapidly draw a sample from the maximum entropy sampling distribution.

2.4. An example

The two proposed algorithms for the splitting are illustrated in Table 1. This table contains the vectors $\boldsymbol{\pi}^-$ and $\boldsymbol{\pi}^+$ computed for a population of $N = 10$ units with strongly dispersed inclusion probabilities $\boldsymbol{\pi}$ whose sum is equal to $\eta = 5.5$.

On this example, with very heterogeneous inclusion probabilities, the algorithms give relatively different results. For more homogeneous inclusion probabilities, it is expected that the results would have been closer. Indeed, output

TABLE 1
Computation of π_k^- and π_k^+ from π_k by means of the πps-method and the maximum entropy method

		πps-method		maximum entropy	
k	π_k	π_k^-	π_k^+	π_k^-	π_k^+
1	0.01	0.0088	0.0112	0.0071	0.0129
2	0.10	0.0876	0.1124	0.0726	0.1274
3	0.40	0.3506	0.4494	0.3167	0.4833
4	0.40	0.3506	0.4494	0.3167	0.4833
5	0.50	0.4382	0.5618	0.4112	0.5888
6	0.60	0.5258	0.6742	0.5156	0.6844
7	0.70	0.6135	0.7865	0.6284	0.7716
8	0.85	0.7449	0.9551	0.8091	0.8909
9	0.95	0.90	1	0.9354	0.9646
10	0.99	0.98	1	0.9870	0.9930
	5.5	5	6	5	6

vectors $\boldsymbol{\pi}^-$ and $\boldsymbol{\pi}^+$ are, with both methods, continuous functions of $\boldsymbol{\pi}$, and for equal inclusion probabilities, both algorithms give exactly the same results.

Each method has its advantages. On one hand, the πps method is easy to implement, requires very few computations, and provides vectors $\boldsymbol{\pi}^-$ and $\boldsymbol{\pi}^+$ that remain proportional to $\boldsymbol{\pi}$ (and thus to the auxiliary information used initially to compute $\boldsymbol{\pi}$) except for some units for which π_k^+ is equal to 1. On the other hand, when all π_k are in $(0, 1)$, the maximum entropy method does not assign values equal to 1 to elements of the vector $\boldsymbol{\pi}^+$ and allows to implement a maximum entropy design. It is, however, computationally intensive, and can comfortably be used on populations of up to only a few thousand units.

3. Second general solution through an augmented population

Instead of splitting the probability vector in two and calculating $\boldsymbol{\pi}^-$ and $\boldsymbol{\pi}^+$, another method to solve the problem consists of extending the population by adding a supplementary phantom unit labeled $N + 1$. This unit receives the inclusion probability

$$\pi_{N+1} = n + 1 - \eta.$$

So, with the added phantom unit, the sum of all inclusion probabilities has the integer value $n + 1$.

Now, a sampling design is obtained by selecting a sample of size $n+1$ from the augmented population, and considering the induced marginal sampling design on the true population U. Thus the real sample size is n if the phantom unit is selected and $n+1$ if the phantom unit is not selected. Ignoring the phantom unit does not affect the inclusion probabilities for units $1, 2, \ldots, N$. Thus the inclusion probabilities π_k are satisfied. Sampling designs obtained through this method are usually different from those obtained through the methods of Section 2. One exception is given in Proposition 3.1.

Proposition 3.1. *If the method of augmented population is applied with a maximum entropy design, then the sampling design is the same as in Section 2.3.*

A proof is given in Appendix. The method of augmented population is thus a simple way to implement the maximum entropy design when the sum of the inclusion probabilities is not integer, with available implementations of fixed size maximum entropy sampling.

4. Estimation

Usually we are interested in estimating the total of a study variable y, which takes a fixed value y_k for unit k. The population total $Y = \sum_{k \in U} y_k$ can be estimated without bias by the well known Horvitz-Thompson estimator

$$\hat{Y}_{HT} = \sum_{k \in \mathbf{s}} \frac{y_k}{\pi_k}. \tag{4.1}$$

TABLE 2
Population of size $N = 5$ used to exemplify the differences between $\hat{Y}_C$ and $\hat{Y}_{HT}$

k	y_k	π_k	π_k^-	π_k^+
1	3	0.250	0.1723284	0.3794526
2	3	0.250	0.1723284	0.3794526
3	5	0.375	0.2726017	0.5456639
4	8	0.625	0.5402228	0.7662954
5	17	0.875	0.8425187	0.9291355
	36	2.375	2	3

If the π_k's do not sum to an integer, and the sample size cannot be fixed, then it may be more efficient to condition on the realized sample size. For the first method (Section 2), this corresponds to conditioning on the outcome of the random choice between π_k^- and π_k^+. If the conditional inclusion probabilities $\pi_k^- = \Pr(k \in \mathbf{s} \,|\mathbf{s} \in \mathcal{S}_n)$ and $\pi_k^+ = \Pr(k \in \mathbf{s} \,|\mathbf{s} \in \mathcal{S}_{n+1})$ have been calculated, then we may use the conditional estimator

$$\hat{Y}_C = \begin{cases} \displaystyle\sum_{k\in\mathbf{s}} \frac{y_k}{\pi_k^-} & \text{if } \mathbf{s} \in \mathcal{S}_n \\ \displaystyle\sum_{k\in\mathbf{s}} \frac{y_k}{\pi_k^+} & \text{if } \mathbf{s} \in \mathcal{S}_{n+1} \end{cases} \tag{4.2}$$

depending on the realized sample size n or $n+1$. Estimator $\hat{Y}_C$ is unbiased and also unbiased conditional on the actual sample size. The Horvitz-Thompson estimator $\hat{Y}_{HT}$ is not unbiased conditional on the sample size. In this situation we should also use a conditional variance estimator adapted to $\hat{Y}_C$. Indeed, we then have that

$$\operatorname{var}\left(\hat{Y}_C\right) = \mathrm{E}\left(\operatorname{var}\left(\hat{Y}_C \middle| \, |s|\right)\right) + \operatorname{var}\left(\mathrm{E}\left(\hat{Y}_C \middle| \, |s|\right)\right), \tag{4.3}$$

However, $\hat{Y}_C$ is unbiased conditional on the size of s. It follows that the last term in Equation 4.3 is null and that any unbiased estimator of the variance of $\hat{Y}_C$ conditional on size is also an unbiased estimator of the variance of $\hat{Y}_C$.

The gain in efficiency by using $\hat{Y}_C$ instead of $\hat{Y}_{HT}$ can be rather substantial. We illustrate this by an example. The population details are given in Table 2.

In Table 2 we have used the maximum entropy design for calculation of π_k^- and π_k^+ and also for sample selection. The full list of samples, their probabilities and the value of the estimators $\hat{Y}_{HT}$ and $\hat{Y}_C$ are given in Table 3. For this example we get that $V(\hat{Y}_{HT}) = 46.14$ and $V(\hat{Y}_C) = 5.01$. For samples of size $n = 2$, $\hat{Y}_{HT}$ is negatively biased and for samples of size $n = 3$, $\hat{Y}_{HT}$ is positively biased.

5. Applications

As a variance reduction technique, we may want to divide the population into rather small non-overlapping strata and make sure that the sample size varies

TABLE 3
The list of all possible samples, their probabilities and the value of the estimators $\hat{Y}_{HT}$ and $\hat{Y}_C$ for the population given in Table 2

s	$p(\mathbf{s})$	$\hat{Y}_{HT}$	$\hat{Y}_C$
(1,2)	0.0056	24.00	34.82
(1,3)	0.0092	25.33	35.75
(1,4)	0.0205	24.80	32.22
(1,5)	0.0724	31.43	37.59
(2,3)	0.0092	25.33	35.75
(2,4)	0.0205	24.80	32.22
(2,5)	0.0724	31.43	37.59
(3,4)	0.0335	26.13	33.15
(3,5)	0.1185	32.76	38.52
(4,5)	0.2633	32.23	34.99
(1,2,3)	0.0025	37.33	24.98
(1,2,4)	0.0056	36.80	26.25
(1,2,5)	0.0199	43.43	34.11
(1,3,4)	0.0092	38.13	27.51
(1,3,5)	0.0326	44.76	35.37
(1,4,5)	0.0724	44.23	36.64
(2,3,4)	0.0092	38.13	27.51
(2,3,5)	0.0326	44.76	35.37
(2,4,5)	0.0724	44.23	36.64
(3,4,5)	0.1185	45.56	37.90

as little as possible within the strata. With the proposed methods, this can easily be achieved. If the inclusion probabilities sum to an integer for the entire population, then it is also possible to coordinate the strata sample sizes to have a fixed overall sample size. The procedure is as follows.

Let there be H strata and let $\eta_h = \sum_{k\in U_h} \pi_k$, $n_h < \eta_h < n_h + 1$, where $\sum_{h=1}^{H} \eta_h = n$ and n is an integer. Some strata should get sample size n_h and some $n_h + 1$. Stratum h should get sample size $n_h + 1$ with probability $q_h = \eta_h - n_h$. As $\sum_{h=1}^{H} q_h = n - \sum_{h=1}^{H} n_h$ is an integer (say m), any fixed size sampling design on $\{1, \ldots, H\}$, with inclusion probabilities $q_1, q_2, \ldots, q_H$, can be used to select m strata that will get sample sizes $n_h + 1$. For each selected stratum we calculate $\boldsymbol{\pi}_h^+$ and for the non-selected strata we calculate $\boldsymbol{\pi}_h^-$. Now we can apply a fixed-size unequal probability design within each domain, with these inclusion probability vectors, to select our sample. By this procedure we respect the initial inclusion probabilities, have a minimum variance in the domain sample sizes, and a fixed overall sample size.

Another important application is that of bootstrap methods for a finite population. Antal and Tillé (2011) have shown that if the sample is selected without replacement with unequal inclusion probabilities, the bootstrap method must take the sampling design into account. They propose a two-phase bootstrap procedure. In the first phase, a set of units is selected once in the bootstrap sample with the same unequal probabilities as the original sample. In the second phase, the units that are not selected in the first phase are resampled with equal probabilities with replacement. However, during the first phase, the sum of the inclusion probabilities in the sample is not integer. For this first phase,

the authors needed to have sampling procedures that can be used when inclusion probabilities do not sum to an integer. The authors used a size-constrained sampling design as described in Section 2.

6. Discussion

Once vectors $\boldsymbol{\pi}^+$ and $\boldsymbol{\pi}^-$ are computed and randomly chosen, any known fixed-size sampling method can be applied on the selected vector, which makes these procedures very general. Both solutions of Section 2 and 3 cover all probability distributions that satisfy Conditions (2.1) and (2.2). However, only a limited number of fixed size sampling procedures are actually implemented. Using these sampling procedures on an augmented population as in Section 3 or after a trial, with updated inclusion probabilities as in Section 2 usually leads to different sampling designs. Maximum entropy sampling, is a notable exception where both methods coincide to the same sampling design.

The advantage of the splitting technique is that π_k^- and π_k^+ are calculated and we may use the estimator $\hat{Y}_C$. The advantage of using the method of augmented population with a phantom unit is that it does not require calculation of π_k^- and π_k^+. Thus it is easier to implement, but if the conditional inclusion probabilities are not calculated, we may not use the more efficient estimator $\hat{Y}_C$. For some designs it is however possible to calculate the conditional inclusion probabilities and use the more efficient estimator even if the sample was selected by the approach with an augmented population.

Acknowledgements

The authors are grateful to a referee and an associate Editor whose valuable comments have helped them improve this manuscript.

Appendix A: Proofs

A.1. Proof of Proposition 2.2

The equation $\sum_{k\in U} \pi_k^- = n$ is automatically satisfied, along with most other properties that immediately follow from the definition of $\boldsymbol{\pi}^-$. It is also easy to see that, in algorithm 2.1, $n+1-|A| \geq \sum_{i\notin A} \pi_i$ and $1 \geq \pi_k^+ \geq \pi_k$, for all $k \in U$. The only part that requires a proof is the assertion that $\boldsymbol{\pi}^-$ is truly an inclusion probability vector, that is to say that all π_k^- lie in $[0, 1]$. This is proved below:

(i) Since $\pi_k^+ \geq \pi_k$, for all k, and $\pi_k = (1-q)\pi_k^- + q\pi_k^+$, we immediately have that $\pi_k^- \leq \pi_k \leq 1$.
(ii) For the other inequality, we need to prove that $\pi_k \geq q\pi_k^+$, for all k. Two cases may occur:

– if $k \in A$ then, by definition, $(n+2-k)\pi_k \geq \sum_{i=k}^{N} \pi_i$ which is, after having subtracted π_k from both sides of the inequality, equivalent to

$$\pi_k \geq \frac{\eta - \sum_{i=1}^{k} \pi_i}{n+1-k}.$$

Hence, it is sufficient to prove that

$$\frac{\eta - \sum_{i=1}^{k} \pi_i}{n+1-k} \geq q = \eta - n. \tag{A.1}$$

– If $k \notin A$ and $|A| < n+1$, $\pi_k^+ = (n+1-|A|)\pi_k / \sum_{i \notin A} \pi_i$ and we need to prove that

$$\frac{\eta - \sum_{i \in A} \pi_i}{n+1-|A|} \geq q = \eta - n. \tag{A.2}$$

In order to prove inequalities (A.1) and (A.2), it is sufficient to prove that

$$\frac{\eta - p}{n+1-p} \geq \frac{\eta - n}{n+1-n},$$

if $p \leq n$, or:

$$\frac{n+1-n}{n+1-p} \geq \frac{\eta - n}{\eta - p},$$

(every term is positive since $p \leq n < \eta < n+1$). But the function

$$f_{n,p} = x \to \frac{x-n}{x-p}$$

is well defined for all $p \leq n$ and $x > p$ and is non-decreasing thanks to the fact that $n \geq p$. Using this property and the fact that $\eta < n+1$ proves the result and even that, if $|A| < n$ or if π_k, $k \in A$ are not all equal to 1, then $\pi_k > 0$ implies that $\pi_k^- > 0$. □

A.2. Proof of Proposition 3.1

Consider the sampling design $p(\cdot)$ on U, obtained by selecting a sample $\tilde{\mathbf{s}}$ with inclusion probabilities

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k, \ldots, \pi_N, 1-q)'$$

on the augmented population with the fixed size maximum entropy design $\tilde{p}(\cdot)$, and retaining the sample $\mathbf{s}$ given by the first N selection indicators. We have that

$$p(\mathbf{s}) = \tilde{p}(\tilde{\mathbf{s}}) = \frac{\exp(\tilde{\boldsymbol{\lambda}}'\tilde{\mathbf{s}})}{\sum_{\tilde{\mathbf{s}} \in \tilde{S}_{n+1}} \exp(\tilde{\boldsymbol{\lambda}}'\tilde{\mathbf{s}})},$$

where $\tilde{S}_{n+1}$ is the set of all samples of size $n+1$ in the augmented population and $\tilde{\boldsymbol{\lambda}}$ is a vector such that the inclusion probabilities are exact (see for example

Chen et al., 1994). Since $\tilde{p}(\cdot)$ has fixed size, such a vector $\tilde{\boldsymbol{\lambda}}$ is defined up to an additive constant. We can thus force $\tilde{\lambda}_{N+1} = 0$. If $\boldsymbol{\lambda}$ is the vector that holds the N first coordinates of $\tilde{\boldsymbol{\lambda}}$, we then have that

$$p(\mathbf{s}) = \frac{\exp(\boldsymbol{\lambda}'\mathbf{s})}{\sum_{\mathbf{s}\in S_n} \exp(\boldsymbol{\lambda}'\mathbf{s}) + \sum_{\mathbf{s}\in S_{n+1}} \exp(\boldsymbol{\lambda}'\mathbf{s})}.$$

Hence, $p(\cdot)$ is the maximum entropy design on $S_n \cup S_{n+1}$ with inclusion probabilities $\pi_1, \ldots, \pi_N$. □

References

Aires, N. (2000). *Techniques to calculate exact inclusion probabilities for conditional Poisson sampling and Pareto πps sampling designs*. Doctoral thesis, Chalmers University of Technology and Göteborg University, Göteborg, Sweden.

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, **106(494)**, 534–543. MR2847968

Bondesson, L. and Grafström, A. (2011). An extension of Sampford's method for unequal probability sampling. *Scandinavian Journal of Statistics*, **38(2)**, 377–392. MR2829606

Berger, Y. G. (1998). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 209–226. MR1624693

Brewer, K. R. W. and Donadio, M. E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, **29**, 189–196.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*. Springer, New York. MR0681289

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Hayward, CA: Institute of Mathematical Statistics. MR0882001

Chen, S. X., Dempster, A. P., and Liu, J. S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, **81**, 457–469. MR1311090

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, **43**, 1470–1480. MR0345337

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes, France.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, **85**, 89–101. MR1627234

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, **35**, 1491–1523. MR0178555

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker. MR0627744

Hardy, G. H., Littlewood, J. E. and Pólya, G. (1956). *Inequalities*. Cambridge Univ. Press.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, **138**, 1428–1432. MR2388021

Sampford, M. R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, **54**, 499–513. MR0223051

Särndal, C. -E. and Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York. MR1140409

Tillé, Y. (2006). *Sampling Algorithms*. Springer, New York. MR2225036