Biometrika (2004), **91**, 4, *pp.* 913–927

Printed in Great Britain

Coordination, combination and extension of balanced samples

By YVES TILLÉ

Groupe de Statistique, Université de Neuchâtel, Espace de l'Europe 4, Case postale 805, 2002 Neuchâtel, Switzerland

yves.tille@unine.ch

and ANNE-CATHERINE FAVRE

Département de hydrologie statistique, Institut national de la recherche scientifique, 2800 rue Einstein, C.P. 7500 Sainte-Foy (Québec), Canada G1V 4C7

anne-catherine_favre@inrs-ete.uquebec.ca

Summary

The cube method allows the selection of balanced samples on several auxiliary variables with equal or unequal inclusion probabilities. Practical implementation of the cube method has raised questions concerning the selection of a multi-phase balanced sampling design, the rebalancing of an unbalanced sampling design by completing it with another sample, the selection of a balanced sample from an unbalanced sample and the coordination of balanced samples. This paper provides a complete solution of all these problems.

Some key words: Cube method; Multi-phase sampling; Rotation; Sample coordination; Unequal probability sampling.

1. Introduction

Balanced sampling is a method of sample selection that preserves a given set of sample inclusion probabilities and satisfies that the Horvitz–Thompson estimators of known auxiliary variables are the same or nearly the same as the corresponding true totals. Interest in balanced sampling was already pointed out more than 50 years ago by Yates (1946). Several partial solutions of the balanced sampling problem have been proposed by Yates (1946), Thionet (1953), Deville et al. (1988), Ardilly (1991), Deville (1992), Hedayat & Majumdar (1995) and Valliant et al. (2000). A general solution, the cube method, that allows the selection of balanced samples on several auxiliary variables, with equal or unequal inclusion probabilities, has been proposed by Deville & Tillé (2004a); see also Tillé (2001, Ch. 8). This method is based on a geometric representation of a sampling design, and preserves exactly a set of given inclusion probabilities.

Since the implementation of the technique by A. Bousabaa, J. Lieber, R. Sirolli and F. Tardieu, it has been applied many times. An SAS macro allows the selection of balanced samples with up to several tens of auxiliary variables and several tens of thousands of population units. Applications have raised a lot of questions which are discussed in this paper.

A first problem is the selection of several nonoverlapping balanced samples from the same population. In some cases, the samples can be selected together, and in some other cases not. A simple examination of the problem reveals unexpected difficulties. We show that, with unequal inclusion probabilities, the complement of a balanced sample is in general not balanced. Nevertheless, we propose a way of balancing simultaneously the sample and its complement. We also show that multi-phase balanced sampling is possible if we modify the auxiliary variables.

Balancing an unbalanced sample using a supplementary sample is the second problem addressed in this paper. The supplementary sample can come from the same population, from another population, such as births, or from a changing population. Along the lines of Kish & Scott (1971), a modification to the auxiliary variables, a suitable choice of the balancing variables and the use of inclusion probability conditional on the first draw provide solutions to this problem. Finally, we show how to select a balanced sample from another sample that is not balanced.

2. NOTATION

The units in the study population are designated by a label $k = 1, \ldots, N$. For a population that changes in time the set of labels of the units is denoted by U_t during a finite set of time points $t = 1, \ldots, T$. The size of U_t is denoted by N_t. As time passes, new units can appear and others disappear. The set of birth labels at time t is given by $U_t \backslash U_{t-1}$, while the set of death labels is $U_{t-1} \backslash U_t$. For simplicity, we assume that each unit has a label k which does not change with time, so that at any time we can identify without ambiguity the units of U_t and pair them with the corresponding units of U_{t+1}. The unit identified by k is not necessarily present in each population U_t $(t = 1, \ldots, T)$.

We also have auxiliary variables x_t^j $(j = 1, \ldots, p)$ which are known for every unit at any time t. As units are born and die, the values of the jth auxiliary variable at time t on unit k are denoted by x_{kt}^j $(j = 1, \ldots, p, k \in U)$; in general, they change with time. The values taken by the variable of interest y_{kt} also evolve. Since the auxiliary variables are assumed to be known for all the population units, the vector of totals,

$$X_t = \sum_{k \in U_t} x_{kt},$$

is known, where $x_{kt} = (x_{kt}^1, \ldots, x_{kt}^j, \ldots, x_{kt}^p)'$. The elements of the vector x_{kt} can be the values of any variables known for the whole population. For instance, if the aim is to select a sample of municipalities, the x-variables might be the area of the municipality, the number of inhabitants, the proportion of foreigners or the number of accommodations. The x-variables can also depend on the inclusion probabilities, or can be a constant, for example $x_{kt}^j = 1$ $(k \in U_t)$, or an indicator variable of a stratum.

The objective is to estimate the total of the variable of interest given by

$$Y_t = \sum_{k \in U_t} y_{kt}.$$

In the paper, we consider for simplicity that y_{kt} is scalar, although the multivariate generalisation follows directly.

A sample s_t is a subset of U_t. Let $p_t(s_t)$ denote the probability of selecting the sample s_t at time t. We denote by S_t the random sample, such that

$$p_t(s_t) = \mathrm{pr}(S_t = s_t),$$

and by $n(S_t)$ the size of sample S_t. The random samples have a joint distribution given by

$$\mathrm{pr}(S_1 = s_1, \ldots, S_t = s_t, \ldots, S_T = s_T) = p(s_1, \ldots, s_t, \ldots, s_T).$$

Further notation is $\pi_{kt} = \text{pr}(k \in S_t)$, the inclusion probability of unit k at time t, for all $k \in U_t$, and $\pi_{k(t-1)t} = \text{pr}(k \in S_{t-1} \cap S_t)$, the inclusion probability of unit k at both instants $t-1$ and t, for all $k \in U_{t-1} \cap U_t$.

At each time t, we consider the Horvitz–Thompson estimators given by

$$\hat{Y}_t = \sum_{k \in S_t} \frac{y_{kt}}{\pi_{kt}}, \quad \hat{X}_t = \sum_{k \in S_t} \frac{x_{kt}}{\pi_{kt}}.$$

Theoretically the joint inclusion probabilities are $\pi_{k\ell t} = \text{pr}(k \text{ and } \ell \in S_t)$, and $\pi_{k\ell(t-1)t} = \text{pr}(k \in S_{t-1} \text{ and } \ell \in S_t)$. Note that the quantity $\pi_{k\ell(t-1)t}$ is not symmetrical in k and ℓ. Indeed, we have

$$\pi_{\ell k(t-1)t} = \text{pr}(\ell \in S_{t-1} \text{ and } k \in S_t).$$

The variances of the Horvitz–Thompson estimator are

$$\text{var}(\hat{Y}_t) = \sum_{k \in U_t} \sum_{\ell \in U} \frac{y_{kt}}{\pi_{kt}} \Delta_{k\ell t} \frac{y_{\ell t}}{\pi_{\ell t}}, \quad \text{var}(\hat{X}_t) = \sum_{k \in U_t} \sum_{\ell \in U_t} \frac{x_{kt}}{\pi_k} \Delta_{k\ell t} \frac{x'_{\ell t}}{\pi_{\ell t}}, \tag{1}$$

where

$$\Delta_{k\ell t} = \pi_{k\ell t} - \pi_{kt}\pi_{\ell t} \quad (k, \ell \in U_t).$$

However, unless these inclusion probabilities can be expressed analytically, for simple random sampling, the second-order inclusion probabilities are of limited interest. We will outline in § 3 that estimation of the variances of the totals is possible in balanced sampling using only the first-order inclusion probabilities π_{kt}. In most of the cases, we will consider only two waves, that is $t = 1, 2$.

3. Balanced sampling

At a given time, the objective is to select a sample with given selection probabilities that is assumed to be of one stage and balanced on the available auxiliary variables x^j_{kt}. A family of algorithms is available (Deville & Tillé, 2004a) for selecting a balanced random sample. It is thus possible to select a sample at time t, so that the identities

$$\hat{X}_t = \sum_{k \in S_t} \frac{x^j_{kt}}{\pi_{kt}} = \sum_{k \in U_t} x^j_{kt} \quad (j = 1, \ldots, p) \tag{2}$$

hold exactly or nearly exactly; since sample sizes are integers one cannot always satisfy (2) exactly (Deville & Tillé, 2004a).

If a sample is balanced then $\hat{X}_t$ is not random. Thus, a necessary and sufficient condition for a sampling design to be balanced is that

$$\text{var}(\hat{X}_t) = \sum_{k \in U_t} \sum_{\ell \in U_t} \frac{x_{kt}}{\pi_{kt}} \Delta_{k\ell t} \frac{x'_{\ell t}}{\pi_{\ell t}} = 0. \tag{3}$$

An approximation of the variance of $\hat{Y}_t$ in (1) of the Horvitz–Thompson estimator has been proposed for a balanced sampling design by Deville & Tillé (2004b). The ideas developed in this paper are the following. Let

$$\hat{Y}_{t,\text{Poiss}} = \sum_{k \in S_{t,\text{Poiss}}} \frac{y_k}{\pi_k}, \quad \hat{X}_{t,\text{Poiss}} = \sum_{k \in S_{t,\text{Poiss}}} \frac{x_k}{\pi_k},$$

where $S_{t,\mathrm{Poiss}}$ is a sample selected by a Poisson sampling design of inclusion probabilities $\breve{\pi}_{kt}$. If we assume that the balanced sampling design maximises or nearly maximises entropy, the variance can be approximated by the variance of a conditional Poisson sampling design, which can be written

$$\mathrm{var}(\hat{Y}_t) = \mathrm{var}(\hat{Y}_{t,\mathrm{Poiss}} | \hat{X}_{t,\mathrm{Poiss}} = X_t).$$

If we suppose that, under Poisson sampling, $(\hat{Y}_{t,\mathrm{Poiss}} \;\; \hat{X}'_{t,\mathrm{Poiss}})'$ approximately has a multinormal distribution, which is asymptotically true, we obtain

$$\begin{aligned}\mathrm{var}(\hat{Y}_{t,\mathrm{Poiss}} | \hat{X}_t = X_t) &= \mathrm{var}[\hat{Y}_{t,\mathrm{Poiss}} + (X_t - \hat{X}_{t,\mathrm{Poiss}})\{\mathrm{var}(\hat{X}_{t,\mathrm{Poiss}})\}^{-1} \mathrm{cov}(\hat{X}_{t,\mathrm{Poiss}}, \hat{Y}_{t,\mathrm{Poiss}})] \\ &= \mathrm{var}(\hat{Y}_{t,\mathrm{Poiss}} - \hat{X}'_{t,\mathrm{Poiss}} B_{t,\mathrm{Poiss}}) \\ &= \sum_{k \in U_t} \frac{(y_{kt} - x'_{kt} B_{t,\mathrm{Poiss}})^2}{\pi_{kt}^2} \breve{\pi}_{kt}(1 - \breve{\pi}_{kt}),\end{aligned}$$

where

$$B_{t,\mathrm{Poiss}} = \left\{ \sum_{k \in U_t} \frac{x_{kt} x'_{kt}}{\pi_{kt}^2} \breve{\pi}_{kt}(1 - \breve{\pi}_{kt}) \right\}^{-1} \sum_{k \in U_t} \frac{x_{kt} y_{kt}}{\pi_{kt}^2} \breve{\pi}_{kt}(1 - \breve{\pi}_{kt}).$$

The $\breve{\pi}_{kt}$'s are the inclusion probabilities of the Poisson sampling design, and are not equal to the π_{kt}'s. The $\breve{\pi}_{kt}$'s cannot be computed exactly, but Deville & Tillé (2004b) have proposed using

$$\breve{\pi}_{kt}(1 - \breve{\pi}_{kt}) = \frac{N}{N - p} \pi_{kt}(1 - \pi_{kt}),$$

which allows us to construct the approximation

$$\frac{N}{N - p} \sum_{k \in U_t} \frac{E_{kt}^2}{\pi_{kt}^2} \pi_{kt}(1 - \pi_{kt}), \tag{4}$$

where $E_{kt} = y_{kt} - x'_{kt} B_{t,\mathrm{Poiss}}$. Deville & Tillé (2004b) used a substantial simulation study to validate this approximation. They also proposed three slightly different approximations for the variance, but approximation (4) has the advantage of depending only on the π_{kt}'s.

4. SELECTION OF SEVERAL NONOVERLAPPING SAMPLES

4·1. *Nonoverlapping samples with unequal probabilities*

Before showing how to coordinate balanced samples, we consider the problem of selecting several nonoverlapping samples with fixed unequal probabilities, first when the samples are selected together and secondly when the samples are selected sequentially at two different times. The difficulties that are already present for the case of unequal probability sampling will become even more complicated for balanced samples.

When the samples must be selected together, Deville & Tillé (2000, p. 219), used Cox's controlled rounding method (Cox, 1987) to show that it is possible to split a population randomly into I parts, with inclusion probabilities $\pi_{k,i}$ ($k \in U, i = 1, \ldots, I$), where

$$\sum_{i=1}^{I} \pi_{k,i} = 1, \quad \sum_{k \in U} \pi_{k,i} = n_i, \quad \sum_{i=1}^{I} n_i = N.$$

Using this method one can select nonoverlapping samples with unequal probabilities. However, the samples must be selected together, not sequentially.

If the samples must be selected sequentially from a population U that does not change with time, the problem becomes more intricate. We first select a sample S_1 with unequal probabilities π_{k1} in U. Suppose that a second sample S_2 with no unit in common with S_1 is needed and must have unconditional inclusion probabilities π_{k2}, where $\pi_{k1} + \pi_{k2} \leqslant 1$, for all $k \in U$. The second sample S_2 must be drawn with conditional probabilities π_{kb} from the complement of S_1, that is $\bar{S}_1 = U \backslash S_1$. Note that, if unit k is selected from the random sample $\bar{S}_1$, we must assume that π_{kb} can be a function of S_1 and thus random. In order to ensure that the global inclusion probabilities π_{k2} are fixed for the second sample, the selection probabilities π_{kb} have to satisfy

$$\pi_{k2} = (1 - \pi_{k1}) E(\pi_{kb}) \quad (k \in U).$$

Two alternative strategies can be explored. In the first strategy we select the sample S_1 according to π_{k1} and then compute

$$\pi_{kb} = \begin{cases} \pi_{k2}/(1 - \pi_{k1}) & (k \notin S_1), \\ 0 & (k \in S_1). \end{cases}$$

A sample S_2 is selected from $U \backslash S_1$ with inclusion probabilities π_{kb}. A remaining problem is that $\sum_{k \in U \backslash S_1} \pi_{kb} = n_2(S_1)$ is a random variable which depends on the sample selected at the first stage. The size of the sample of the second wave is then necessarily random. In the second strategy we select the sample S_1 according to π_{k1}; then we compute

$$\tilde{\pi}_{kb|S_1} = \begin{cases} \dfrac{n_2 \{\pi_{k2}/(1 - \pi_{k1})\}}{\sum_{\ell \in U \backslash S_1} \{\pi_{\ell 2}/(1 - \pi_{\ell 1})\}} & (k \notin S_1), \\ 0 & (k \in S_1), \end{cases} \tag{5}$$

where $n_2 = \sum_{k \in U} \pi_{k2}$ is the fixed size of the sample. In this case $\tilde{\pi}_{kb|S_1}$ is a random variable depending on S_1, but the sample has a fixed size. The inclusion probabilities of the second sample are not exactly π_{k2} but can be expressed as

$$\begin{aligned} E\{(1 - \pi_{k1}) \tilde{\pi}_{kb|S_1}\} &= E\left[(1 - \pi_{k1}) \frac{n_2 \{\pi_{k2}/(1 - \pi_{k1})\}}{\sum_{\ell \in U \backslash S_1} \{\pi_{\ell 2}/(1 - \pi_{\ell 1})\}} \right] \\ &= n_2 \pi_{k2} E\left[\frac{1}{\sum_{\ell \in U \backslash S_1} \{\pi_{\ell 2}/(1 - \pi_{\ell 1})\}} \right]. \end{aligned}$$

Each method has a disadvantage which we can easily circumvent by selecting the sample of the first wave S_1 so that $\sum_{U \backslash S_1} \pi_{kb}$ is fixed; that is

$$\sum_{k \in U \backslash S_1} \pi_{kb} = E\left(\sum_{k \in U \backslash S_1} \pi_{kb} \right) = \sum_{k \in U} (1 - \pi_{k1}) \pi_{kb}, \tag{6}$$

which can be expressed as

$$\sum_{k \in U} \pi_{kb} - \sum_{k \in S_1} \pi_{kb} = \sum_{k \in U} (1 - \pi_{k1}) \pi_{kb},$$

or

$$\sum_{k \in S_1} \pi_{kb} = \sum_{k \in U} \pi_{kb} \pi_{k1},$$

or

$$\sum_{k \in S_1} \frac{\pi_{k1} \pi_{kb}}{\pi_{k1}} = \sum_{k \in U} \pi_{kb} \pi_{k1}.$$

The last expression amounts to selecting a first-wave sample S_1 balanced on the variable $x_k = \pi_{k1} \pi_{kb}$. The inclusion probabilities are thus unchanged.

If the balancing condition (6) is verified, then we combine the advantages of both methods. Indeed the size of the second sample is fixed and the probabilities of the second wave are not random. However this result is limited because, if the inclusion probabilities of the second sample are known during the first sample as is assumed in (6), then we can simply select both samples by means of the method described in Deville & Tillé (2000). This simple application shows that using fixed-size multi-phase sampling techniques in the case of unequal probabilities is much more complex than for simple random sampling.

4·2. *Nonoverlapping balanced samples*

The selection of several balanced samples with unequal probabilities is even more difficult. First, it is easy to show that, if a sample S is balanced on the variables $x_1 \ldots, x_p$, then its complement $U \backslash S$ is not necessarily balanced.

PROPOSITION 1. *The complement of S is balanced on $x_1, \ldots, x_p$ if and only if S is balanced on $\pi_k x_k/(1 - \pi_k)$.*

Proof. The random sample S and its complement $U \backslash S$ have the same variance-covariance operators, namely $\Delta = (\Delta_{k\ell})$, where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$. As we have seen in (3), a sample S is balanced if

$$\sum_{k \in U} \sum_{\ell \in U} \frac{x_k}{\pi_k} \Delta_{k\ell} \frac{x'_\ell}{\pi_\ell} = 0.$$

As $\mathrm{pr}(k \in U \backslash S) = 1 - \pi_k$, the sample $U \backslash S$ is balanced if

$$\sum_{k \in U} \sum_{\ell \in U} \frac{x_k}{1 - \pi_k} \Delta_{k\ell} \frac{x'_\ell}{1 - \pi_\ell} = 0.$$

The proof follows directly from these two expressions of variances. □

The following corollary is obvious.

COROLLARY 1. *If S is selected with equal probabilities, then $U \backslash S$ is balanced on the same variables as S.*

Proposition 1 illustrates the limits of the notion of balanced samples. However the method works well in the case of sampling with equal probabilities and has been successfully implemented for the redeveloped census of the INSEE.

The problem becomes even more awkward when we want to split a population into q samples $S_1, \ldots, S_q$, which do not overlap, with unequal probabilities $\pi_{k,i}$, for $i = 1, \ldots, q$ and $k = 1, \ldots, N$.

Proposition 2. *If two nonoverlapping samples S_i and S_j are selected with inclusion probabilities $\pi_{k,i}$ and $\pi_{k,j}$, respectively, and are both balanced on x_k, then their union is not necessarily balanced on x_k.*

Proposition 2 follows from the nonlinearity of the Horvitz–Thompson estimator. Indeed, if

$$\hat{Y}_i = \sum_{k \in S_i} \frac{y_k}{\pi_{k,i}}, \quad \hat{Y}_j = \sum_{k \in S_j} \frac{y_k}{\pi_{k,j}}, \quad \hat{Y}_{ij} = \sum_{k \in S_i \cup S_j} \frac{y_k}{\pi_{k,i} + \pi_{k,j}},$$

then $\hat{Y}_{ij}$ is not a linear combination of $\hat{Y}_i$ and $\hat{Y}_j$. Thus, the balancing property is lost by the reunion of the samples. Nevertheless, the balancing properties remain unchanged when the designs have equal inclusion probabilities. Moreover, it is easy to show that, if $\pi_{k,i} \propto \pi_{k,j}$ for all $k \in U$, then the union of two balanced samples is balanced.

4·3. *Multi-phase balanced sampling*

It is possible to select a balanced sample S_2 from a balanced sample S_1 in such a way that the subsample remains balanced on the same variables. If S_1 is a balanced sample on the variables $x_1, \ldots, x_p$, with probabilities π_{k1} then

$$\sum_{k \in S_1} \frac{x_k}{\pi_{k1}} = \sum_{k \in U} x_k.$$

If

$$z_k = x_k / \pi_{k1}, \tag{7}$$

and if we select a sample S_2 from S_1 that is balanced on the variables z_k with inclusion probabilities π_{k2} then we have

$$\sum_{k \in S_2} \frac{z_k}{\pi_{k2}} = \sum_{k \in S_1} z_k = \sum_{k \in U} x_k.$$

Multi-phase sampling is thus possible while keeping the balancing property.

5. Rebalancing using a sample from another population

5·1. *Conditional balanced design with conditional inclusion probabilities*

Initially this problem arose in the French redeveloped census. In the larger municipalities five nonoverlapping samples of addresses must be selected. These samples are balanced on demographic variables, and each one of them will be used for a year. The French census therefore has five ongoing samples and all of them must remain balanced.

The question is how to incorporate the new constructions into the rotation groups with minimal distortion to equilibrium. The set of new constructions can be considered as a new population, a population of births, in which five new samples must be selected to complete the existing samples. A first question is therefore how to select a sample in a new population in order that the union of the old and new samples remains balanced.

If only one sample must be selected in the old and new populations, the problem can be formalised in the following way. Consider two nonoverlapping populations U_1 and U_2 of sizes N_1 and N_2 respectively. A random sample S_1 has been drawn from U_1 with inclusion probabilities π_{k1} ($k \in U_1$). The sample S_1 is in general not balanced because of the drift of the sample caused by the deaths and the evolution of the balancing variables. The Horvitz–Thompson estimator for the auxiliary variables,

$$\hat{X}_1 = \sum_{k \in S_1} \frac{x_k}{\pi_{k1}},$$

is therefore not equal to the population total $X_1 = \sum_{k \in U_1} x_k$. The problem consists of selecting a random sample S_2 from U_2 with given inclusion probabilities π_{k2}. In the French census, for instance, all the π_{k2} are equal to $\frac{1}{5}$, but in another problem π_{k2} can be computed so as to optimise the variance. Sample S_2 must be selected in such a way that

$$\hat{X}_1 + \hat{X}_2 = X_1 + X_2, \tag{8}$$

where

$$\hat{X}_2 = \sum_{k \in S_2} \frac{x_k}{\pi_{k2}}, \quad X_2 = \sum_{k \in U_2} x_k.$$

In other words, we want to choose S_2 so as to rebalance S_1. In order for relationship (8) to be realised, sample S_2 must satisfy the balancing equation

$$\hat{X}_2 = T(S_1), \tag{9}$$

where $T(S_1) = X_1 + X_2 - \hat{X}_1$. Note that S_2 must be balanced on a random value $T(S_1)$ which depends on the S_1 selected from U_1. In order to satisfy equality (9), S_2 will be selected with conditional inclusion probabilities $\tilde{\pi}_{k2|S_1}$ that depend on S_1 and not on S_2. In order to extract the probabilities $\tilde{\pi}_{k2|S_1}$ we note first that, if we compute the conditional expectation of equation (9) conditional on S_1, we obtain $E(\hat{X}_2|S_1) = T(S_1)$. Thus the $\tilde{\pi}_{k2|S_1}$'s must satisfy

$$E(\hat{X}_2|S_1) = T(S_1), \quad E(\tilde{\pi}_{k2|S_1}) = \pi_{k2},$$

which can be written

$$\sum_{k \in U_2} \frac{x_k}{\pi_{k2}} \tilde{\pi}_{k2|S_1} = T(S_1), \quad E(\tilde{\pi}_{k2|S_1}) = \pi_{k2}. \tag{10}$$

A solution can be found by looking for the $\tilde{\pi}_{k2|S_1}$'s that minimise the chi-squared distance

$$\sum_{k \in U_2} \frac{(\tilde{\pi}_{k2|S_1} - \pi_{k2})^2}{\pi_{k2} w_k},$$

under the constraints

$$\sum_{k \in U_2} \frac{x_k \tilde{\pi}_{k2|S_1}}{\pi_{k2}} = T(S_1).$$

The w_k's are weights that can be chosen arbitrarily. The solution of this optimisation problem is given by

$$\tilde{\pi}_{k2|S_1} = \pi_{k2} + \{T(S_1) - X_2\}' \left(\sum_{\ell \in U_2} \frac{x_\ell x_\ell' w_\ell}{\pi_{\ell 2}^2} \right)^{-1} \frac{x_k w_k}{\pi_{k2}}. \tag{11}$$

It is easy to see that equation (11) satisfies (10). Unfortunately, in some cases, equation (11) may yield probabilities $\tilde{\pi}_{k2|S_1} > 1$ or $\tilde{\pi}_{k2|S_1} < 0$. Nevertheless, this problem can be avoided with a good choice of the w_k's. For instance, by taking $w_k = \pi_{k2}(1 - \pi_{k2})$, the difference between $\tilde{\pi}_{k2|S_1}$ and π_{k2} will be small when π_{k2} is close to 0 or 1.

Example 1. An illustrative example is the case where the only auxiliary variable is $x_k = \pi_{k1}$ if $k \in U_1$ and $x_k = \pi_{k2}$ if $k \in U_2$, which corresponds to the fixed sample size problem. In this case, if $n(S_1)$ is the size of S_1, and $n(S_2)$ is the size of S_2, we obtain

$$\tilde{\pi}_{k2|S_1} = \pi_{k2} + [E\{n(S_1)\} - n(S_1)] \frac{w_k}{\sum_{k \in U_2} w_k}. \tag{12}$$

Thus, if $w_k = 1$, we obtain

$$\tilde{\pi}_{k2|S_1} = \pi_{k2} + \frac{E\{n(S_1)\} - n(S_1)}{N_2},$$

and, if $w_k = \pi_{k2}(1 - \pi_{k2})$,

$$\tilde{\pi}_{k2|S_1} = \pi_{k2} \times \left(1 + \frac{1 - \pi_{k2}}{\sum_{k \in U_2} \pi_{k2}(1 - \pi_{k2})} \times [E\{n(S_1)\} - n(S_1)] \right).$$

In either case, the π_{k2}'s are adjusted so that

$$\sum_{k \in U_2} \tilde{\pi}_{k2|S_1} = E\{n(S_2)\} + E\{n(S_1)\} - n(S_1).$$

In practice, the problem can be solved using a single balanced sample selection program. The sample must be selected with inclusion probabilities $\tilde{\pi}_{k2|S_1}$ and the balancing variables must be redefined as

$$z_k = x_k \tilde{\pi}_{k2|S_1} / \pi_{k2}.$$

Indeed, the balancing equations,

$$\sum_{k \in S_2} \frac{z_k}{\tilde{\pi}_{k2|S_1}} = \sum_{k \in U_2} z_k,$$

imply (9) and thus $S_1 \cup S_2$ is a balanced sample from $U_1 \cup U_2$.

Note that the problem of the French census is more complex, because five non-overlapping samples must be selected together in the old and new populations. However this operation is not too difficult because the units are selected with equal inclusion probabilities. In this case, the complement of a balanced design is also balanced. The five new samples can thus be selected successively in such a way that the balancing conditions are satisfied.

5·2. *Variance of a rebalanced sample*

The variance of a rebalanced sample can be deduced by using the same methodology as for the variance of a two-phase sampling design (Särndal & Swensson, 1987). The variance of the total estimator $\hat{Y}$ is calculated by conditioning on S_1:

$$\begin{aligned}\operatorname{var}(\hat{Y}) &= \operatorname{var}(\hat{Y}_1 + \hat{Y}_2)\\ &= E\{\operatorname{var}(\hat{Y}_1 + \hat{Y}_2|S_1)\} + \operatorname{var}\{E(\hat{Y}_1 + \hat{Y}_2|S_1)\}\\ &= E\{\operatorname{var}(\hat{Y}_2|S_1)\} + \operatorname{var}\{\hat{Y}_1 + E(\hat{Y}_2|S_1)\}.\end{aligned} \tag{13}$$

The first term of (13) is an expectation of the variance under balanced sampling. We can thus use the same methodology as for expression (4). The variance is approximated by the variance of a conditional Poisson sampling design, which can be written

$$\operatorname{var}(\hat{Y}_2|S_1) = \operatorname{var}\{\hat{Y}_{2,\mathrm{Poiss}}|S_1, \hat{X}_{2,\mathrm{Poiss}} = T(S_1)\},$$

where $\hat{Y}_{2,\mathrm{Poiss}}$ and $\hat{X}_{2,\mathrm{Poiss}}$ are the estimators when the sample is selected by means of a Poisson sampling. Again, if we suppose that, under Poisson sampling, the vector $(\hat{Y}_{2,\mathrm{Poiss}}, \hat{X}'_{2,\mathrm{Poiss}})'$ approximately has a multinormal distribution, we obtain

$$\begin{aligned}&E[\operatorname{var}\{\hat{Y}_{2,\mathrm{Poiss}}|S_1, \hat{X}_{2,\mathrm{Poiss}} = T(S_1)\}]\\ &\quad= E(\operatorname{var}[\hat{Y}_{2,\mathrm{Poiss}} + \{T(S_1) - \hat{X}_{2,\mathrm{Poiss}}\}\{\operatorname{var}(\hat{X}_{2,\mathrm{Poiss}})\}^{-1}\operatorname{cov}(\hat{X}_{2,\mathrm{Poiss}}, \hat{Y}_{2,\mathrm{Poiss}})|S_1])\\ &\quad= E[\operatorname{var}\{\hat{Y}_{2,\mathrm{Poiss}} + (X_1 + X_2 - \hat{X}_1 - \hat{X}_{2,\mathrm{Poiss}})'B_{2,\mathrm{Poiss}}|S_1\}]\\ &\quad= E\{\operatorname{var}(\hat{Y}_{2,\mathrm{Poiss}} - \hat{X}'_{2,\mathrm{Poiss}}B_{2,\mathrm{Poiss}}|S_1)\}\\ &\quad\doteq \frac{N}{N-p}E\left\{\sum_{k\in U_2}\frac{(y_k - x'_kB_{2,\mathrm{Poiss}})^2}{\pi_k^2}\tilde{\pi}_{k2|S_1}(1-\tilde{\pi}_{k2|S_1})\right\},\end{aligned}$$

where

$$B_{2,\mathrm{Poiss}} = \left\{\sum_{k\in U_2}\frac{x_kx'_k}{\pi_k^2}\tilde{\pi}_{k2|S_1}(1-\tilde{\pi}_{k2|S_1})\right\}^{-1}\sum_{k\in U_2}\frac{x_ky_k}{\pi_k^2}\tilde{\pi}_{k2|S_1}(1-\tilde{\pi}_{k2|S_1}).$$

The expectation of the second term of (13) is

$$\begin{aligned}E(\hat{Y}_2|S_1) &= \sum_{k\in U_2}\frac{y_k}{\pi_{k2}}E(I_k|S_1) = \sum_{k\in U_2}\frac{y_k}{\pi_{k2}}\tilde{\pi}_{k2|S_1}\\ &= \sum_{k\in U_2}\frac{y_k}{\pi_{k2}}\left[\pi_{k2} + \{T(S_1) - X_2\}'\left(\sum_{\ell\in U_2}\frac{x_\ell x'_\ell w_\ell}{\pi_{\ell 2}^2}\right)^{-1}\frac{x_kw_k}{\pi_{k2}}\right]\\ &= Y_2 + (X_1 - \hat{X}_1)'B_2,\end{aligned}$$

where

$$B_2 = \left(\sum_{\ell\in U_2}\frac{x_\ell x'_\ell w_\ell}{\pi_{\ell 2}^2}\right)^{-1}\sum_{k\in U_2}\frac{x_ky_kw_k}{\pi_{k2}^2}.$$

We then obtain

$$
\begin{aligned}
\operatorname{var}\{\hat{Y}_1 + E(\hat{Y}_2 | S_1)\} &= \operatorname{var}\{\hat{Y}_1 + Y_2 + (X_1 - \hat{X}_1)' B_2\} \\
&= \operatorname{var}(\hat{Y}_1 - \hat{X}_1' B_2) \\
&= \operatorname{var}\left(\sum_{k \in S_1} \frac{y_k - x_k' B_2}{\pi_{k1}} \right).
\end{aligned}
$$

Finally, the variance is

$$
\operatorname{var}(\hat{Y}) = \frac{N}{N-p} E\left\{ \sum_{k \in U_2} \frac{(y_k - x_k' B_{2,\mathrm{Poiss}})^2}{\pi_k^2} \tilde{\pi}_{k2|S_1}(1 - \tilde{\pi}_{k2|S_1}) \right\} + \operatorname{var}\left(\sum_{k \in S_1} \frac{y_k - x_k' B_2}{\pi_{k1}} \right). \tag{14}
$$

The second term of (14) depends on S_1 for which the sampling design was not specified. The expression of variance works with any sampling design for the selection of S_1. Equation (14) shows that the variance of $\operatorname{var}(\hat{Y})$ can be expressed as a variance of residuals. The regression coefficients B_2 and $B_{2,\mathrm{Poiss}}$ are slightly different but are both computed from U_2.

Based on estimators for $B_{2,\mathrm{Poiss}}$ and B_2, the estimator of the variance is

$$
\widehat{\operatorname{var}}(\hat{Y}) = \frac{n}{n-p} \sum_{k \in S_2} \frac{(y_k - x_k' \hat{B}_{2,\mathrm{Poiss}})^2}{\pi_k^2} (1 - \tilde{\pi}_{k2|S_1}) + \widehat{\operatorname{var}}\left(\sum_{k \in S_1} \frac{y_k - x_k' \hat{B}_2}{\pi_{k1}} \right),
$$

where the second term is constructed by means of an estimator of the variance under the sampling scheme yielding S_1.

6. Rebalancing using a sample from the same population

This problem has been posed with respect to the INSEE master sample. In each region, the primary units, i.e. sets of municipalities, are selected with unequal inclusion probabilities proportional to the size by using a sampling design balanced on demographic and economic variables. Some regions have requested a supplementary sample in order to make survey extensions. The problem is thus that of finding out if it is possible to supplement the sample by a new nonoverlapping sample in such a way that the union of the two samples remains balanced.

Formally, the problem consists of selecting a nonoverlapping sample in a population in which a sample has already been selected. Suppose that a sample S_1, not necessarily balanced, has been selected in U, with inclusion probabilities π_{k1}. Note that, by Proposition 1, even if S_1 is balanced, in general $U \backslash S_1$ is not. The aim is thus to supplement S_1 with a nonoverlapping sample S_2 in such a way that

$$
\operatorname{pr}\{k \in (S_1 \cup S_2)\} = \pi_k,
$$

for all $k \in U$, and the completed sample is balanced, that is

$$
\sum_{k \in (S_1 \cup S_2)} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k. \tag{15}
$$

The probabilities π_k are given either by implementation of an optimisation criterion or from practical considerations. For instance, the sampling design used to select the master sample of the INSEE is a self-weighting multi-stage design that determines the values of the π_k's.

Since $(S_1 \cap S_2) = \varnothing$, $\pi_{k2} = \text{pr}(k \in S_2) = \pi_k - \pi_{k1}$, for $k \in U$. To obtain (15), we must have

$$\sum_{k \in S_2} \frac{x_k}{\pi_k} = T(S_1), \tag{16}$$

where

$$T(S_1) = \sum_{k \in U} x_k - \sum_{k \in S_1} \frac{x_k}{\pi_k}.$$

In order to satisfy (16), we can select a balanced sample S_2 from $U \backslash S_1$ with conditional inclusion probabilities $\tilde{\pi}_{kb|S_1}$ and balancing variables $z_k = x_k \tilde{\pi}_{kb|S_1}/\pi_k$. The probabilities $\tilde{\pi}_{kb|S_1}$ are defined as

$$\tilde{\pi}_{kb|S_1} = \begin{cases} \pi_{kb} + \{T(S_1) - V(S_1)\}' \left(\sum_{\ell \in U \backslash S_1} \dfrac{x_\ell x_\ell' w_\ell}{\pi_\ell^2} \right)^{-1} \dfrac{x_k w_k}{\pi_k} & (k \notin S_1), \\ 0 & (k \in S_1), \end{cases} \tag{17}$$

where

$$\pi_{kb} = \begin{cases} \pi_{k2}/(1 - \pi_{k1}), & \text{if } k \notin S_1, \\ 0, & \text{if } k \in S_1, \end{cases}$$

$$V(S_1) = \sum_{k \in U \backslash S_1} \frac{x_k \pi_{kb}}{\pi_k},$$

and the w_k's are weights that can be chosen arbitrarily. Again we recommend the use of $w_k = \pi_{kb}(1 - \pi_{kb})$, which should allow us to avoid $\tilde{\pi}_{kb|S_1} < 0$ and $\tilde{\pi}_{kb|S_1} > 1$. Indeed, when π_k is close to 0 or 1, the weights w_k will be very small and, from equation (17), we can see that π_{kb} will be close to $\tilde{\pi}_{kb}$.

With the conditional inclusion probabilities given in (17), we obtain

$$E\left(\sum_{k \in S_2} \frac{x_k}{\pi_k} \middle| S_1 \right) = \sum_{k \in U \backslash S_1} \frac{x_k}{\pi_k} \tilde{\pi}_{kb|S_1} = \sum_{k \in U \backslash S_1} \frac{x_k}{\pi_k} \pi_{kb} + \{T(S_1) - V(S_1)\} = T(S_1).$$

Now, let us compute the expectation of $\tilde{\pi}_{kb}$. First compute

$$Q(S_1) = T(S_1) - V(S_1) = \sum_{k \in S_1} x_k \frac{1 - \pi_k}{(1 - \pi_{k1})\pi_k} - \sum_{k \in U} x_k \frac{\pi_{k1}(1 - \pi_k)}{(1 - \pi_{k1})\pi_k}.$$

Note that $E\{Q(S_1)\} = 0$, and that $Q(S_1)$ is a single-stage Horvitz–Thompson estimator centred in S_1. A reasonable assumption for one-stage sampling designs is that

$$\frac{Q(S_1)}{N} = O_p\left[\sqrt{\left\{ \frac{N - n(S_1)}{N n(S_1)} \right\}} \right],$$

where $O_p(1/a)$ is a quantity that remains bounded in probability when multiplied by a. Moreover, if

$$W = \left(\sum_{\ell \in U \backslash S_1} \frac{x_\ell x_\ell' w_\ell}{\pi_\ell^2} \right)^{-1}, \tag{18}$$

another reasonable assumption is that

$$NW - E(NW) = O_p\left[\sqrt{\left\{ \bar{w} \frac{N - n(U \backslash S_1)}{Nn(U \backslash S_1)} \right\}} \right] = O_p\left(\bar{w} \sqrt{\left[\frac{n(S_1)}{N\{N - n(S_1)\}} \right]} \right),$$

where $\bar{w} = N^{-1} \sum_{k \in U} w_k$. Since $E\{Q(S_1)\} = 0$, we have $E\{Q(S_1)W\} = E[Q(S_1)\{W - E(W)\}]$. Now, by (17) and (18), we have, by assuming that $nw_k/(\bar{w}\pi_k N)$ is close to 1,

$$\begin{aligned} E(\tilde{\pi}_{kb|S_1}) &= \pi_{k2} + E\left\{ Q(S_1) W \frac{x_k w_k}{\pi_k \bar{w}} \right\} = \pi_{k2} + \frac{N x_k}{n} E\left[Q(S_1)\{W - E(W)\} \frac{n w_k}{N \pi_k \bar{w}} \right] \\ &= \pi_{k2} + \frac{N x_k}{n} O_p\left(\frac{1}{N} \right) = \pi_{k2} + E\left\{ O_p\left(\frac{x_k}{n} \right) \right\}. \end{aligned} \tag{19}$$

Thus, under very mild regularity conditions, we have that $E(\tilde{\pi}_{kb|S_1}) \simeq \pi_{k2}$. In practice, the problem can be solved by redefining the auxiliary variables. A balanced sample S_2 is thus selected from $U \backslash S_1$ with inclusion probabilities $\tilde{\pi}_{kb|S_1}$. The balancing variables are

$$z_k = \frac{x_k}{\pi_k} \tilde{\pi}_{kb|S_1}.$$

Indeed, $\sum_{k \in S_2} (z_k / \tilde{\pi}_{kb|S_1}) = \sum_{k \in U} z_k$, implying (16).

7. Sample coordination

Another important problem is that of sample coordination, and especially of coordinating balanced samples. The aim is to select a balanced sample that does not overlap with a sample, or a set of samples, already selected. Formally, the problem is the following. A sample S_1 has already been selected from a population U with inclusion probabilities π_{k1}. The aim is to select a balanced nonoverlapping sample S_2 from $U \backslash S_1$ with inclusion probabilities π_{k2}. The balancing equations are thus

$$\sum_{k \in S_2} \frac{x_k}{\pi_{k2}} = X.$$

First suppose that $\pi_{k1} + \pi_{k2} \leqslant 1$, for all $k \in U$. The problem is that S_2 must be selected from $U \backslash S_1$, and that, even if S_1 is balanced, $U \backslash S_1$ is not necessarily balanced. Sample S_2

will be selected from $U\backslash S_1$ by the use of conditional inclusion probabilities $\pi_{kb|S_1}$ defined as

$$\tilde{\pi}_{kb|S_1} = \begin{cases} \pi_{kb} + \{X - V(S_1)\}' \left(\sum_{\ell \in U\backslash S_1} \frac{x_\ell x_{\ell'} w_\ell}{\pi_{\ell 2}^2} \right)^{-1} \frac{x_k w_k}{\pi_{k2}} & (k \notin S_1), \\ 0 & (k \in S_1), \end{cases} \tag{20}$$

where

$$\pi_{kb} = \begin{cases} \pi_{k2}/(1-\pi_{k1}), & \text{if } k \notin S_1, \\ 0, & \text{if } k \in S_1, \end{cases}$$

$$V(S_1) = \sum_{k \in U\backslash S_1} \frac{x_k \pi_{kb}}{\pi_{k2}},$$

and the w_k's are weights that can be chosen arbitrarily. We always recommend $w_k = \pi_{kb}(1-\pi_{kb})$. In practice, the problem can be solved by selecting the sample S_2 from $U\backslash S_1$ with the inclusion probabilities $\tilde{\pi}_{kb|S_1}$, and the balancing variables

$$z_k = \frac{x_k}{\pi_{k2}} \tilde{\pi}_{kb|S_1}.$$

Example 2. An already treated application is the case where $x_k = \pi_{k2}$ and $w_k = \pi_{kb}$. We obtain

$$z_k = \tilde{\pi}_{kb|S_1} = \frac{\pi_{kb} \sum_{\ell \in U} \pi_{\ell 2}}{\sum_{\ell \in U\backslash S_1} \pi_{\ell b}},$$

which is the solution proposed in equation (5) for sampling with unequal probabilities.

In a real coordination problem, we can have $\pi_{k1} + \pi_{k2} > 1$, for some $k \in U$. The aim is then to select a sample S_2 as disconnected as possible from S_1. The following solution can be applied. Define the conditional inclusion probabilities as follows:

$$\tilde{\pi}_{kb|S_1} = \begin{cases} 1 & (k \notin S_1 \text{ and such that } \pi_{k1} + \pi_{k2} \geqslant 1), \\ \pi_{kb} + \{X - V(S_1)\}' \left(\sum_{\ell \in A} \frac{x_\ell x'_\ell w_\ell}{\pi_{\ell 2}^2} \right)^{-1} \frac{x_k w_k}{\pi_{k2}} & (k \in A), \\ 0 & (k \in S_1 \text{ and such that } \pi_{k1} + \pi_{k2} < 1), \end{cases}$$

where

$$A = \{[k|(\pi_{k1} + \pi_{k2}) \geqslant 1 \text{ and } k \in S_1] \cup [k|(\pi_{k1} + \pi_{k2}) < 1 \text{ and } k \notin S_1]\},$$

$$\pi_{kb} = \begin{cases} 1, & \text{if } \pi_{k1} + \pi_{k2} \geqslant 1 \text{ and } k \notin S_1, \\ (\pi_{k1} + \pi_{k2} - 1)/\pi_{k1}, & \text{if } \pi_{k1} + \pi_{k2} \geqslant 1 \text{ and } k \in S_1, \\ \pi_{k2}/(1-\pi_{k1}), & \text{if } \pi_{k1} + \pi_{k2} < 1 \text{ and } k \notin S_1, \\ 0 & \text{if } \pi_{k1} + \pi_{k2} < 1 \text{ and } k \in S_1, \end{cases}$$

$$V(S_1) = \sum_{k \in U} \frac{x_k \pi_{kb}}{\pi_{k2}}.$$

Next define new variables $z_k = x_k \tilde{\pi}_{kb|S_1}/\pi_{k2}$. The sample S_2 is selected with inclusion probabilities $\tilde{\pi}_{kb|S_1}$ balanced on z_k.

Acknowledgement

The authors thank Jean Dumais and Jean-Claude Deville for their helpful comments and Ines Pasini for her help in typing the paper. The authors are also grateful to an associate editor and two anonymous referees for their constructive suggestions. This research was supported by a grant of the Swiss Federal Office of Statistics.

References

Ardilly, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. *Ann. Écon. Statist.* **23**, 91–113.

Cox, L. (1987). A constructive procedure for unbiased controlled rounding. *J. Am. Statist. Assoc.* **82**, 520–4.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop Auxiliary Information in Surveys*, pp. 21–40. Örebro, Sweden: Statistics Sweden.

Deville, J.-C. & Tillé, Y. (2000). Selection of several unequal probability samples from the same population. *J. Statist. Plan. Infer.* **86**, 215–27.

Deville, J.-C. & Tillé, Y. (2004a). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893–912.

Deville, J.-C. & Tillé, Y. (2004b). Variance approximation under balanced sampling. *J. Statist. Plan. Infer.* To appear.

Deville, J.-C., Grosbras, J.-M. & Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT, Proceedings in Computational Statistics*, Ed. D. G. Edwards and N. E. Kaun, pp. 255–66. Heidelberg: Physica Verlag.

Hedayat, A. & Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *J. Statist. Plan. Infer.* **44**, 237–47.

Kish, L. & Scott, A. (1971). Retaining units after changing strata and probabilities. *J. Am. Statist. Assoc.* **66**, 461–70.

Särndal, C.-E. & Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *Int. Statist. Rev.* **55**, 279–94.

Thionet, P. (1953). *La Théorie des Sondages*. Paris: INSEE, Imprimerie Nationale.

Tillé, Y. (2001). *Théorie des Sondages: Échantillonnage et Estimation en Populations Finies*. Paris: Dunod.

Valliant, R., Dorfman, A. & Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys (with Discussion). *J. R. Statist. Soc.* A **109**, 12–43.

[*Received January* 2002. *Revised August* 2003]