Published in Statistics & Probability Letters 74, issue 1, 31-37, 2005
which should be used for any reference to this work

Optimal allocation in balanced sampling

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
Chaire en hydrologie statistique
Institut National de la Recherche Scientifique, Eau Terre et Environnement
490 de la Couronne, Québec (QC) G1K 9A9, Canada
anne-catherine_favre@ete.inrs.ca

March 14, 2005

SUMMARY

The development of new sampling methods allows the selection of large balanced samples. In this paper we propose a method for computing optimal inclusion probabilities for balanced samples. Next, we show that the optimal Neyman allocation is a particular case of this method.

Some key-words: balanced sampling, optimal inclusion probabilities, variance minimization, Neyman's allocation.

1 Introduction

The interest of balanced sampling was already pointed out more than 50 years ago by Yates (1946). Several partial solutions of balanced sampling methods have been proposed by Yates (1946), Thionet (1953), Deville et al. (1988), Ardilly (1991), Deville (1992), Hedayat and Majumdar (1995) and Valliant et al. (2000). A general solution (the cube method) allowing the selection of balanced samples on several auxiliary variables, with equal or unequal inclusion probabilities, has been proposed by Deville and Tillé (2005a,b) (on the cube method, see also Tillé, 2001, chap. 8). Since the implementation of the technique by Bousabaa et al. (1999) and Tardieu (2001), a multitude of applications have been realized. A SAS macro allows the selection of balanced samples with up to several tens of auxiliary variables and several tens of thousands of population units. In this paper, we determine the optimal inclusion probabilities that minimize the variance of the Horvitz-Thompson estimator using balanced samples.

The paper is organized as follows. The notation is defined in Section 2. Section 3 is devoted to the definition of balanced sampling. In Section 4, we give a new method

to compute optimal inclusion probabilities for balanced sampling. The application to stratification developed in Section 5 shows that this method is linked with Neyman's optimal allocation. Section 6 draws the main conclusions.

2 Notation

The general problem consists in studying a population i.e. a set of statistical units, (households, establishments). Each unit can be designed by a label $k = 1, \ldots, N$. The set of labels of the units is denoted U_t during a finite set of time points $t = 1, \ldots, T$. The size of U_t is denoted N_t. This population changes with time. New units can appear and others disappear. The set of births labels at time t is given by $U_t \backslash U_{t-1}$, while the set of death labels is $U_{t-1} \backslash U_t$. Also assume that each unit has a label k which does not change with time. We can thus at any time 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 $U_t, t = 1, \ldots, T$.

In this population, we also have auxiliary variables $x_t^j, j = 1, ..., p$, which are known at any time t. As units are born and die, the values of the jth auxiliary variable at time t on unit k is denoted $x_{kt}^j, j = 1, ..., p, k \in U$, and, in general, 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

$$\mathbf{X}_t = \sum_{k \in U_t} \mathbf{x}_{kt},$$

is known, where

$$\mathbf{x}_{kt} = (x_{kt}^1, \ldots, x_{kt}^j, \ldots, x_{kt}^p)'.$$

The elements of vector $\mathbf{x}_{kt}$ can be the values of any variable known on the whole population. For instance, if the aim is to select a sample of municipalities, the x-variables can 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 ($x_{kt}^j = 1, k \in U_t$), or an indicator variables of a stratum.

The objective is to estimate a function of the variables of interest y_{kt} given by

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

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

$$p_t(s_t) = Pr\left(S_t = s_t\right),$$

and by $n(S_t)$ the size of sample S_t. So, we denote by $\pi_{kt} = Pr(k \in S_t)$ the inclusion probability of unit k at time t, for all $k \in U_t$; $\pi_{k(t-1)t} = Pr\left[k \in S_{t-1} \cap S_t\right]$ the inclusion probability of unit k at both instants $t-1$ and t, for all $k \in U_{t-1} \cap U_t$; $\pi_{k\ell t} = Pr\left[k \text{ and } \ell \in S_t\right]$ the joint inclusion probability.

3 Balanced sampling

At each time t, we consider the Horvitz-Thompson estimator given by

$$\widehat{Y}_t = \sum_{k \in S_t} \frac{y_{kt}}{\pi_{kt}}.$$

The variance of this estimator is

$$\text{var}(\widehat{Y}_t) = \sum_{k \in U_t} \sum_{\ell \in U_t} \frac{y_{kt}}{\pi_{kt}} \Delta_{k\ell t} \frac{y_{\ell t}}{\pi_{\ell t}}, \tag{1}$$

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

At a given time, the objective is to select a sample using a given sampling design, which is assumed to be of one stage, and balanced on the available auxiliary variables x_{kt}^j and with unequal probabilities. A family of algorithms is available (Deville and Tillé, 2005a) to select a balanced random sample. It is thus possible to select at time t a sample, so that the identities

$$\widehat{\mathbf{X}}_t = \sum_{k \in S_t} \frac{x_{kt}^j}{\pi_{kt}} = \sum_{k \in U_t} x_{kt}^j, j = 1, ..., p,$$

are verified exactly or nearly exactly. If a sample is balanced then $\widehat{\mathbf{X}}_t$ is not random.

With a balanced design, a very good approximation of the variance given in (1), using only the first order inclusion probabilities, has been given in Deville and Tillé (2005b) and has been validated by a set of simulations:

$$\text{var}(\widehat{Y}_t) \cong \frac{N}{N-p} \sum_{k \in U_t} \frac{E_{kt}^2}{\pi_{kt}^2} \pi_{kt} \left(1 - \pi_{kt}\right), \tag{2}$$

where

$$E_{kt} = y_{kt} - \mathbf{x}_{kt}\mathbf{B}_t,$$
$$\mathbf{x}_{kt} = (x_{kt}^1, ..., x_{kt}^j, ..., x_{kt}^p)'$$

and

$$\mathbf{B}_t = \left\{ \sum_{k \in U_t} \frac{\mathbf{x}_{kt}\mathbf{x}_{kt}'}{\pi_{kt}^2} \pi_{kt} \left\{1 - \pi_{kt}\right\} \right\}^{-1} \sum_{k \in U_t} \frac{\mathbf{x}_{kt} y_{kt}}{\pi_{kt}^2} \pi_{kt} \left\{1 - \pi_{kt}\right\}.$$

Deville and Tillé (2005b) have proposed three other slightly different approximations for the variance, however equation (2) has the advantage of depending only on π_{kt}.

4 Estimation of the optimal inclusion probabilities

In most of cases the inclusion probabilities are fixed. When the populations must be partitioned into rotation groups the inclusion probabilities must be equal. If a self-weighted multi-stage sampling design is used, the inclusion probabilities must be proportional to a measure of size of the primary units. However, when the choice of the inclusion probabilities is free, we could look for optimal inclusion probabilities in the sense where the variance of the variable of interest is minimized.

Suppose that we want to determine optimal inclusion probabilities for time $t = 2$. The optimal balanced sampling design is obtained by minimizing the variance (2) under the constraint that the expected sample size be $\sum_{k \in S_2} \pi_{k2} = n_2$. After some algebra, we get the optimal inclusion probabilities

$$\pi_{k2}^{opt} \propto |E_{k2}|, (k \in U), \tag{3}$$

with the following constraints

$$0 \leq \pi_{k2} \leq 1, (k \in U),$$

and

$$\sum_{k \in U_2} \pi_{k2} = n_2,$$

where n_2 is the fixed size of S_2.

However, E_{k2} depends on π_{k2}. Expression (3) must thus be viewed as a system of equations to solve in order to compute the optimal inclusion probabilities. A fixed-point algorithm could be used to solve this system of equations.

Nevertheless, a solution cannot be directly found because (even in stratification) $E_{k2}, k \in U_2$, and the $y_{k2}, k \in U_2$ are never known. A piece of the optimization will thus be lost. Only $\mathbf{x}_{k2}$ can be available in some cases on the whole population. We then have to estimate the $|E_{k2}|$'s by means of a model whose parameters are determined using a former survey at time point $t = 1$. Using this survey, we could find an approximation of optimal inclusion probabilities using

$$\widehat{\pi}_{k1}^{opt} \propto |\widehat{E}_{k1}|, \tag{4}$$

where $|\widehat{E}_{k1}|$ is computed from the estimator of the residuals $\widehat{E}_{k1}$ with

$$\widehat{E}_{k1} = y_{k1} - \mathbf{x}_{k1}\widehat{\mathbf{B}}_1,$$

where

$$\widehat{\mathbf{B}}_1 = \left\{ \sum_{k \in S_1} \frac{\mathbf{x}_{k1}\mathbf{x}'_{k1}}{\pi_{k1}^2} \frac{\pi_{k1}(1 - \pi_{k1})}{\pi_{k1}} \right\} \times \sum_{k \in S_1} \frac{\mathbf{x}_{k1} y_{k1}}{\pi_{k1}^2} \frac{\pi_{k1}(1 - \pi_{k1})}{\pi_{k1}},$$

and π_{k1} are the inclusion probabilities at time $t = 1$.

Nevertheless, the aim is to compute an estimation of $\widehat{E}_{k2}$ in order to get an approximation of the optimal inclusion probabilities for time point $t = 2$. To estimate the $|E_{k2}|$,'s a linear model can be used at time $t = 1$

$$|\widehat{E}_{k1}|^2 = \sum_{j=1}^{p} \beta_j x_{k1}^j + \epsilon_k, \tag{5}$$

where the ϵ_k's are residuals. The β_j's can be estimated in minimizing in β_j the quantity

$$\sum_{k \in S_1} \frac{1}{\pi_{k1}} \left(|E_{k1}|^2 - \sum_{j=1}^{p} \beta_j x_{k1}^j \right)^2.$$

Finally we obtain a predictive tool for $|E_{k2}|$ which is given by

$$|\widehat{E}_{k2}|^2 = \sum_{j=1}^{p} \widehat{\beta}_j x_{k2}^j,$$

where $\widehat{\beta}_j$ is the estimator of β_j.

The optimal inclusion probabilities are estimated by

$$\widehat{\pi}_{k2}^{opt} = n \frac{|\widehat{E}_{k2}|}{\sum_{k \in U} |\widehat{E}_{k2}|}. \tag{6}$$

For units for which the quantities computed in (6) are larger than 1, we set $\widehat{\pi}_{k2}^{opt} = 1$. Next, the quantities are recalculated using (6) restricted to the remaining units, as it is generally done for the computation of unequal inclusion probabilities.

5 Application to the optimal allocation in stratification

A stratified sampling design is a particular case of balanced sampling design. Let U_{ht} be the h^{th} stratum at time t, with $h = 1, ..., H$, N_{ht} the size of the h^{th} stratum at time t, and n_{ht} the sample size in the h^{th} stratum at time t. The auxiliary variables are

$$x_{kt}^h = \begin{cases} 1 & \text{if } k \in U_{ht} \\ 0 & \text{if } k \notin U_{ht} \end{cases} \quad (k \in U, h = 1, \ldots, H).$$

Equation (3) becomes

$$\pi_{k2}^{opt} \propto |E_{k2}| = |y_{kt} - \overline{Y}_{h(k)2}|, (k \in U), \tag{7}$$

where $\overline{Y}_{h(k)2}$ is the population mean of the variable of interest y computed in the stratum of unit k at time 2. Equation (4) can be expressed as

$$\widehat{\pi}_{k1}^{opt} \propto |\widehat{E}_{k1}| = y_k - \overline{y}_{h(k)1}, \tag{8}$$

where $\overline{y}_{h(k)1}$ is the sample mean of the variable of interest y computed in the stratum of unit k at time 1. After some algebra, the least square estimator of β_j obtained by minimizing (5) is given by

$$\widehat{\beta}_h = \frac{1}{n_h} \sum_{k \in S_1 \cap U_{h1}} |\widehat{E}_{k1}|^2 = s_{h1}^2,$$

where s_{h1}^2 is the sample variance computed in stratum h at time t :

$$s_{ht}^2 = \frac{1}{n_h} \sum_{k \in S_t} (y_{kt} - \bar{y}_h)^2,$$

and

$$\bar{y}_h \frac{1}{n_h} \sum_{k \in S_t} y_{kt}.$$

Finally, Equation (6) can be written as

$$\widehat{\pi}_{k2}^{opt} = n\frac{|\widehat{E}_{k2}|}{\sum_{k\in U}|\widehat{E}_{k2}|} = n\frac{s_{h(k)1}}{\sum_{\ell\in U} s_{h(\ell)1}} = n\frac{s_{h(k)1}}{\sum_{h=1}^{H} N_{h2}s_{h1}}.$$

Units of the same stratum have thus the same probability to be selected. The estimated optimal size of the stratum sample is thus

$$\widehat{n}_{hopt} = \sum_{k\in U_h} n_2\frac{s_{h(k)1}}{\sum_{j=1}^{H} N_{j2}s_{j1}} = n_2\frac{N_{h2}s_{h1}}{\sum_{j=1}^{H} N_{j2}s_{j1}},$$

where $s_{h(k)1}$ is the standard deviation of variable y_{k1} in the stratum of unit k at time 1.

The method proposed in the previous section consists in this case of selecting a stratified sample with the optimal Neyman (1934) allocation, where the variance of the stratum are estimated by means of a previous sample. The notion of optimal inclusion probabilities is thus related to the idea of optimization used by Neyman for the stratified random design. Neyman has proposed the optimal stratum sizes and therefore optimal inclusion probabilities. In his well known paper, Neyman makes the assumption that the inclusion probabilities of the same stratum are equal.

6 Discussion

Application of the optimal allocation to stratification, a special case of balanced sampling, highlights the link with Neyman's optimization. The proposed technique should allow one to take into account the size effect present in business surveys without exaggerating this effect. If several variables of interest are available, we can compute the optimal probabilities for each of them and average these probabilities. In repeated surveys, we can also smooth the inclusion probabilities with respect to time, which would probably improve the possibilities of coverage in sample coordination problems, because the differences would then be less steep. Moreover it is always sound to increase the inclusion probabilities according to the estimation of the non-response rate using the previous surveys.

References

Ardilly, P. (1991). Echantillonnage représentatif optimum à probabilités inégales. *Annales d'Economie et de Statistique*, 23:91–113.

Bousabaa, A., Lieber, J., and Sirolli, R. (1999). La macro cube. Technical report, INSEE, Rennes.

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*, Örebro (Sweden).

Deville, J.-C., Grosbras, J.-M., and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In Verlag, P., editor, *COMPSTAT, Proceedings in Computational Statistics*, pages 255–266.

REFERENCES

Deville, J.-C. and Tillé, Y. (2005a). Efficient balanced sampling: the cube method. *Biometrika, in revision.*

Deville, J.-C. and Tillé, Y. (2005b). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128(2):569–591.

Hedayat, A. and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44:237–247.

Neyman, J. (1934). On the two different aspects of representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Tardieu, F. (2001). Echantillonnage équilibré: de la théorie à la pratique. Technical report, INSEE.

Thionet, P. (1953). *La théorie des sondages.* INSEE, Imprimerie nationale, Paris.

Tillé, Y. (2001). *Théorie des sondages: échantillonnage et estimation en populations finies.* Dunod, Paris.

Valliant, R., Dorfman, A., and Royall, R. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* Wiley Series in Probability and Statistics: Survey Methodology Section, New York.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109:12–43.