Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Stratified balanced sampling

by Guillaume Chauvet

June 2009

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2009
Vol. 35, No. 1, pp. 115-119
Statistics Canada, Catalogue No. 12-001-X

Stratified balanced sampling

Guillaume Chauvet [1]

Abstract

In the selection of a sample, a current practice is to define a sampling design stratified on subpopulations. This reduces the variance of the Horvitz-Thompson estimator in comparison with direct sampling if the strata are highly homogeneous with respect to the variable of interest. If auxiliary variables are available for each individual, sampling can be improved through balanced sampling within each stratum, and the Horvitz-Thompson estimator will be more precise if the auxiliary variables are strongly correlated with the variable of interest. However, if the sample allocation is small in some strata, balanced sampling will be only very approximate. In this paper, we propose a method of selecting a sample that is balanced across the entire population while maintaining a fixed allocation within each stratum. We show that in the important special case of size-2 sampling in each stratum, the precision of the Horvitz-Thompson estimator is improved if the variable of interest is well explained by balancing variables over the entire population. An application to rotational sampling is also presented.

Key Words: Rotational sampling; Maximum entropy; Cube method; Stratification; Unequal probability sampling.

1. Introduction

In the case of stratified sampling, a population U is partitioned into H subpopulations U_h, $h = 1, ..., H$ called strata, in which samples S_h, $h = 1, ..., H$ are selected according to independent sampling designs p_h, $h = 1, ..., H$, respectively. The inclusion probability of unit k is the probability π_k that unit k is in the sample, and the joint inclusion probability is the probability π_{kl} that two distinct units k and l are jointly in the sample. We will write $\boldsymbol{\pi} = (\pi_k)_{k\in U}$ and $\boldsymbol{\pi}^{\mathbf{h}} = (\pi_k)_{k\in U_h}$. We assume that within each stratum U_h, design $p_h(.)$ is of fixed size. In particular, then, we have $\sum_{k\in U_h}\pi_k = n_h$, $h = 1, ..., H$, where n_h denotes the allocation in stratum U_h. In the rest of the paper, we assume that all sample sizes for stratum n_h are integers.

The Horvitz-Thompson estimator $\hat{t}_{\mathbf{z}\pi} = \sum_{k\in S}\mathbf{z}_k/\pi_k = \sum_{h=1}^{H}\hat{t}^h_{\mathbf{z}\pi}$, where $\hat{t}^h_{\mathbf{z}\pi} = \sum_{k\in S_h}\mathbf{z}_k/\pi_k$, provides an unbiased estimate of $t_{\mathbf{z}} = \sum_{k=1}^{H} t^h_{\mathbf{z}}$, where $t^h_{\mathbf{z}} = \sum_{k\in U_h}\mathbf{z}_k$ denotes the total of the variable (vector) $\mathbf{z}$ over U_h. In the particular case where $\mathbf{z}_k = y_k$ is scalar, the variance of the Horvitz-Thompson estimator is given by the Sen-Yates-Grundy variance formula:

$$\mathrm{Var}(\hat{t}_{y\pi}) = \frac{1}{2}\sum_{h=1}^{H}\sum_{k\neq l\in U_h}(\pi_k\pi_l - \pi_{kl})\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2. \qquad (1)$$

This variance is small if the strata are homogeneous with respect to the variable of interest, specifically if y_k/π_k is approximately constant within each stratum.

If a vector $\mathbf{x} = (x_1, ..., x_q)$ of q auxiliary variables is available prior to sample selection for each individual in the population, the sampling within each stratum can be improved with the cube algorithm (Deville and Tillé 2004), which selects balanced samples. Sampling design $p_h(.)$ is said to be balanced on the $\mathbf{x}$ variables if the equations

$$\hat{t}^h_{\mathbf{x}\pi} = t^h_{\mathbf{x}} \qquad (2)$$

are exactly satisfied. The variance of the Horvitz-Thompson estimator is therefore zero for the estimate of the total of the balancing variables. In the particular case where $\mathbf{x} = \boldsymbol{\pi}$, *i.e.*, if the inclusion probability is the only balancing variable, (2) reduces to

$$\sum_{k\in S_h} 1 = \sum_{k\in U_h}\pi_k = n_h. \qquad (3)$$

Hence, stratified sampling of fixed size in each stratum is a particular case of balanced sampling. For any given number of constraints, an exactly balanced sample generally cannot be found. Suppose, for example, that population U_h contains 100 individuals on whom is defined a variable x with two possible values, 0 and 1, and that 53 individuals in the population have the value 0 for that variable. Selecting a size-10 equal-probability sample balanced on variable x would mean selecting a sample containing 5,3 individuals for whom $x = 0$ and 4,7 individuals for whom $x = 1$, which is impossible. Consequently, the goal is generally to select an approximately balanced sample, such that

$$\hat{t}^h_{\mathbf{x}\pi} \simeq t^h_{\mathbf{x}}. \qquad (4)$$

With the cube method (Deville and Tillé 2004), we can select approximately balanced samples on any number of variables, maintaining exactly a predetermined set of inclusion probabilities $\boldsymbol{\pi}$. The method is composed of two phases: the flight phase and the landing phase. At each step in the flight phase, we decide at random to either select or permanently discard one of the population units. At the end of the flight phase, we have, in each stratum U_h, a vector

1. Guillaume Chauvet, Laboratoire de Statistique d'Enquête, CREST/ENSAI, rue Blaise Pascal, Campus de Ker Lann, 35 170 Bruz, France. E-mail: chauvet@ensai.fr.

$\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k\in U_h} \in [0, 1]^N$ that satisfies the following conditions:

$$E(\boldsymbol{\pi}^{\mathbf{h}*}) = \boldsymbol{\pi}^{\mathbf{h}}, \tag{5}$$

$$\sum_{k\in U_h} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U_h} \mathbf{x}_k, \tag{6}$$

$$\text{Card}\{k \in U_h; 0 < \pi_k^* < 1\} \leq q, \tag{7}$$

where E denotes the expectation for the sampling method used in the flight phase. The vector $\boldsymbol{\pi}^{\mathbf{h}*}$ gives the outcome of the flight phase: π_k^* is 1 if unit k is selected, 0 if it is rejected, and between 0 and 1 only if the decision has not been made for unit k after the flight phase. Equations (5) and (6) ensure that the inclusion probabilities and balancing constraints are maintained perfectly at the end of the flight phase. Equation (7) ensures that a decision remains to be made for no more than q individuals in each stratum U_h, where q is the number of balancing variables. The flight phase ends when the balancing constraints can no longer be exactly satisfied. The landing phase consists in defining, conditionally on the outcome of the flight phase, an optimal sampling design defined on the remaining population V. This design is optimal in that it makes it possible to complete the sampling while minimizing the variance, conditionally on the outcome of the flight phase, of the Horvitz-Thompson estimator of the balancing variables. The remaining units are sampled, conditionally on the outcome of the flight phase, with inclusion probabilities $(\pi_k^*)_{k\in V}$, so that the units' unconditional inclusion probabilities $(\pi_k)_{k\in V}$ are maintained exactly.

The measure of entropy associated with a sampling design $p(.)$ defined on population U is given by

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

with the convention $0\log(0) = 0$. Deville and Tillé (2005) have shown that the balanced design with maximum entropy compared with other sampling designs balanced on the same variables and with the same inclusion probabilities can be regarded as the conditional of a Poisson design. Assuming the asymptotic normality of a multivariate Horvitz-Thompson estimator in the case of a Poisson design, they derived a variance approximation formula for the Horvitz-Thompson estimator for a balanced sampling design. In the case of stratified balanced sampling, we have

$$\text{Var}(\hat{t}_{y\pi}) \simeq \sum_{h=1}^{H}\sum_{k\in U_h} \frac{b_k}{\pi_k^2}(y_k - \beta_{\mathbf{h}}\mathbf{x}_k)^2 \tag{8}$$

where $\beta_{\mathbf{h}} = (\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ \mathbf{x}_l'/\pi_l)^{-1}\sum_{l\in U_h} b_l\ \mathbf{x}_l/\pi_l\ y_l/\pi_l$. Deville and Tillé (2005) offer several approximations for the b_k. The simplest is $b_k = \pi_k(1 - \pi_k)$. The variance of the Horvitz-Thompson estimator will be small if, in each stratum, variable of interest y is well explained by balancing variables $\mathbf{x}$.

Sampling will be balanced in each stratum if the number of balancing variables remains small relative to the sample size. In some cases, however, the allocation to each stratum is too small for balanced sampling: if the stratification of the population is very granular, a current practice is to select a size-2 sample in each stratum. In that case, the only condition that can be imposed is a fixed sample size in each stratum.

In the next section, we propose an algorithm based on the cube method that ensures balanced sampling across the entire population for selected variables and exactly maintains the desired allocation within each stratum. Hence, the samples are no longer selected independently in each stratum. Precision is improved in comparison with stratified sampling with fixed sample size in each stratum if the balancing variables are strongly correlated with the variable of interest across the entire population. The algorithm also has the advantage of ensuring approximate balancing in each stratum, and the larger the sample size allocated to the stratum, the more balanced the sampling will be.

2. Stratified balanced sampling with pooling of landing phases

If sample S is selected from U in accordance with the stratified balanced sampling procedure described in section 1, sampling will be balanced in each stratum as long as the landing phase affects a small number of individuals relative to the sample size. Specifically, equation (7) shows that the number of balancing variables must be small relative to the sample allocation in each stratum. In some cases, that constraint cannot be satisfied. The population is often partitioned into very small groups to make the results more relevant, which means decreasing the sample selected in each stratum; the limit generally used is a size-2 sample, which produces an unbiased variance estimator.

Again, we take the case of a population U divided into H strata $U_1, ..., U_H$, for which a vector $\mathbf{x}_k = (\pi_k, \mathbf{z}_k')'$ of auxiliary variables is known. We assume that the variable π_k is one of the balancing constraints, to ensure fixed-size sampling. Where the allocation to each stratum is too small for balanced sampling to apply constraints other than fixed size in each stratum, algorithm 1 provides an alternative sampling method. A flight phase is carried out independently in each of the H strata: we write $\boldsymbol{\pi}^{\mathbf{h}*} = (\pi_k^*)_{k\in U_h}$, $h = 1, ... H$ for the probability vectors obtained at the end of those flight phases, $\boldsymbol{\pi}^* = (\pi_k^*)_{k\in V}$, where V denotes the units that have not yet been sampled or rejected,

and $\mathbf{x}_k^* = (\pi_k^* 1_{k\in U_1}, ..., \pi_k^* 1_{k\in U_H}, \mathbf{z}_k', \pi_k^*/\pi_k)'$. The probability vector obtained after a final flight phase over the set of remaining units is written $\boldsymbol{\pi}^{**} = (\pi_k^{**})_{k\in V}$. The set of units in stratum U_h that have not yet been sampled or rejected at the end of this new flight phase is denoted W_h.

Algorithm 1: Stratified balanced sampling with pooling of landing phases

Step 1. Carry out a flight phase, with balancing variables $\mathbf{x}_k$ and inclusion probabilities π_k, independently in each stratum U_h.

Step 2. Carry out a flight phase, with balancing variables $\mathbf{x}_k^*$ and inclusion probabilities π_k^*, on the set V of units remaining at the end of step 1.

Step 3. Select a fixed-size sample from each subpopulation W_h, with inclusion probabilities π_k^{**}.

The algorithm is based on a method used by the Institut National de la Statistique et des Études Économiques (INSEE) to select the primary units of the 1999 Master Sample. The Master Sample is a sample of dwellings selected in the 1999 Census for use as a sample frame for household surveys. A detailed description of the sampling design for the Master Sample is provided in Bourdalle, Christine and Wilms (2000). The dwellings are first grouped into urban units and rural units. In the subpopulation of units with fewer than 100,000 residents, a sample of about 6% is selected. We have four auxiliary variables (taxable net income and three age groups). The expected number of sample units is too small for stratified sampling by region, with balanced sampling on the four variables in each region. The regions were therefore grouped into eight super-regions, and the sampling processes were coordinated in such a way as to ensure both overall balanced sampling for the four auxiliary variables in each super-region and a fixed sample size in each region.

A similar method was proposed by Rousseau and Tardieu (2004) for the selection of balanced samples from large frames using the CUBE macro available on INSEE's Web site. The macro's run time is approximately proportional to the square of the population size. Note that Chauvet and Tillé (2006) proposed a fast method of balanced sampling whose run time depends only on the size of the population and which can select balanced samples directly from very large populations. The algorithm was programmed into an SAS macro (see Chauvet and Tillé, 2005) and is also available in the R Sampling Package prepared by Matei and Tillé (2006). In both programs, the second flight phase is performed by adding a constraint associated with each stratum to balancing variables $\mathbf{x}_k^*$ and maintaining the fixed-size condition in each stratum.

Using inclusion probabilities vector $\boldsymbol{\pi}^*$ conditionally on the outcome of step 1 ensures that inclusion probabilities vector $\boldsymbol{\pi}$ is maintained by deconditioning from the outcome of step 1. At the end of step 1, equation (6) implies that

$$\forall h = 1...H \qquad \sum_{k\in U_h/0<\pi_k^*<1} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U_h} \mathbf{x}_k - \sum_{k\in U_h/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k},$$

and summing these expressions yields

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^* = \sum_{k\in U} \mathbf{x}_k - \sum_{k\in U/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k}.$$

At the end of step 2, equation (6) leads to

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} = \sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^*,$$

and combining the last two expressions, we get

$$\sum_{k\in V} \frac{\mathbf{x}_k}{\pi_k}\pi_k^{**} + \sum_{k\in U/\pi_k^*=1} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k, \qquad (9)$$

which ensures that balanced sampling on the variables $\mathbf{x}_k$ is maintained exactly at the end of step 2. Step 3 completes the sampling process while maintaining the fixed-size constraint within each stratum U_h and can be carried out by means of a linear program to limit the lack of balance (see Deville and Tillé 2004).

The variance can be approximated with the variance formula proposed by Deville and Tillé (2005), if each flight phase in algorithm 1 is carried out with high entropy. Entropy can be increased substantially by performing a random sort on the population prior to sampling. In this case, the balancing variables are both the $\mathbf{z}_k$ variables and the variables given by the product of the inclusion probabilities and the stratum membership indicators, which ensure a fixed sample size in each stratum. We have

$$\mathrm{Var}(\hat{t}_{y\pi}) \simeq \sum_{k\in U} \frac{b_k}{\pi_k^2}(y_k - \gamma'\mathbf{a}_k)^2 \qquad (10)$$

with $\mathbf{a}_k = (\pi_k 1_{k\in U_1}, ..., \pi_k 1_{k\in U_H}, \mathbf{z}_k')'$ and

$$\gamma = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l}\frac{\mathbf{a}_l'}{\pi_l}\right)^{-1} \sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l}\frac{y_l}{\pi_l}.$$

We can use the variance estimator

$$v(\hat{t}_{y\pi}) = \sum_{k\in S} \frac{b_k}{\pi_k^3}(y_k - \hat{\gamma}'\mathbf{a}_k)^2 \qquad (11)$$

proposed by Deville and Tillé (2005, page 578), with

$$\hat{\gamma} = \left(\sum_{l\in S} \frac{b_l}{\pi_l}\frac{\mathbf{a}_l}{\pi_l}\frac{\mathbf{a}_l'}{\pi_l}\right)^{-1} \sum_{l\in S} \frac{b_l}{\pi_l}\frac{\mathbf{a}_l}{\pi_l}\frac{y_l}{\pi_l}.$$

As shown in the variance approximation formula (10), it is important to note that the independence of the samples

from the various strata is lost with the proposed stratified balanced sampling method. The samples from strata U_1, ..., U_H are coordinated to ensure overall balance across the whole population, which strips them of their independence. The Horvitz-Thompson estimator $\hat{t}_{y\pi}^h$ of total t_{yh} remains unbiased. Its approximate variance is derived from equation (10) by replacing y_k with $y_k \, 1_{k\in U_h}$, and is given by

$$\mathrm{Var}(\hat{t}_{y\pi}^h) \approx \sum_{k\in U} \frac{b_k}{\pi_k^2} (y_k \, 1_{k\in U_h} - (\gamma^h)' \mathbf{a}_k)^2 \qquad (12)$$

with $\mathbf{a}_k = (\pi_k \, 1_{k\in U_1}, \ldots, \pi_k \, 1_{k\in U_H}, \mathbf{z}_k')'$ and

$$\gamma^h = \left(\sum_{l\in U} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{\mathbf{a}_l'}{\pi_l} \right)^{-1} \sum_{l\in U_h} b_l \frac{\mathbf{a}_l}{\pi_l} \frac{y_l}{\pi_l}.$$

In the particular case where the inference does not apply to the entire population but to a domain D that consists of a small number of strata, balanced sampling overall on the $\mathbf{z}$ variables will be of little benefit. The variance of the Horvitz-Thompson estimator $\hat{t}_{y\pi}^D$ of total t_y^D for variable y for that domain will be close to the variance for stratified sampling, which is given by equation (1).

3. Quantitative results

In this section, we carry out a brief simulation study to test the performance of our sampling algorithm. First, we generate a finite population of 1,000, partitioned into 25 strata of equal size containing four variables: two variables of interest, y_1 and y_2; and two auxiliary variables, x_1 and x_2. Variables x_1 and x_2 are generated with a gamma distribution with parameters 4 and 25. Variable y_1 is generated within stratum U_h using the model

$$y_1 = \alpha_{1h} + \varepsilon_h. \qquad (13)$$

The ε_h are generated with a normal distribution with mean 0 and variance σ_h^2. The model used to generate the values of y_1 is given by (13), with $\alpha_{1h} = 20\,h$ and variance σ_h^2 selected to produce a coefficient of determination R^2 approximately equal to 0.60 in each stratum. Variable y_2 is generated with the model

$$y_2 = \alpha_2 + \beta_2 \, x_1 + \gamma_2 \, x_2 + \eta. \qquad (14)$$

The η are generated with a normal distribution with mean 0 and variance ρ^2. The model used to generate the values of y_2 is given by (14), with $\alpha_2 = 500$, $\beta_2 = \gamma_2 = 5$, and variance ρ^2 selected to produce a coefficient of determination R^2 approximately equal to 0,60.

We are interested in estimating the total of variables y_1 and y_2. We select a sample of $n = 25$ ($n = 50$ respectively) units with equal probabilities using three sampling designs:

Design 1: Stratified simple random sampling in each stratum

Design 2: Sampling balanced on variables π, x_1 and x_2

Design 3: Stratified sampling balanced on variables π, x_1 and x_2, with pooling of the landing phases

In the case of stratified sampling, we have an allocation of size 1 (2 respectively) in each stratum. In the balanced designs, each flight phase is preceded by a random sort of the population. The variance associated with design 1 is calculated directly. The variance associated with designs 2 and 3 is approximated on the basis of 10,000 simulations. The results are presented in Table 1.

Table 1
Variance associated with the estimate of the total of two variables for a stratified design, a balanced design and a stratified balanced design with pooling of landing phases

	$n = 25$		$n = 50$	
	Total var.	Total var.	Total var.	Total var.
	y_1	y_2	y_1	y_2
Method	($\times 10^8$)	($\times 10^9$)	($\times 10^8$)	($\times 10^9$)
Design 1	6.05	7.13	2.95	3.48
Design 2	14.31	3.05	7.02	1.40
Design 3	6.00	3.63	2.98	1.54

In each case, the proposed sampling design is comparable with the better of the two strategies. If the variable of interest is approximately constant across all strata, the proposed algorithm produces the same results as the stratified design. If the balancing variables are highly explanatory, the results produced by our algorithm and by direct balanced sampling are equivalent. The slight loss of precision comes from the landing phase: in the case of direct balanced sampling, we attempt to complete the sampling while limiting the lack of balance. With the proposed algorithm, the selected solution is suboptimal because we are imposing the additional constraint of a fixed size in each stratum.

In the case of stratified balanced sampling with pooling of the landing phases, Table 2 shows the variance given by 10,000 simulations and the variance given by the approximation formula (10).

Table 2
Comparison of the variance given by 10,000 simulations and the variance given by the approximation formula in the case of the estimation of two totals for a stratified balanced sampling design with pooling of landing phases

	$n = 25$		$n = 50$	
	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)	Total y_1 ($\times 10^8$)	Total y_2 ($\times 10^9$)
Simulation var.	6.0	3.6	3.0	1.5
Approximated var.	5.9	2.7	2.9	1.3

The approximation formula proposed by Deville and Tillé (2005) is close to exact if the variance associated with the landing phase is small relative to the variance associated with the flight phase. In the case of the y_2 variable, the balancing variables are highly explanatory. The variance is therefore larger for the landing phase than for the flight phase, and the approximation formula understates the actual variance. The variance associated with the landing phase will be considered in future studies.

Acknowledgements

The author is grateful to the referees and an associate editor for their constructive comments and suggestions.

References

Bourdalle, G., Christine, M. and Wilms, L. (2000). Échantillons maître et emploi. Série INSEE Méthodes, Paris, France, 21, 139-173.

Chauvet, G., and Tillé, Y. (2005). New SAS macros for balanced sampling. INSEE, Journées de Méthodologie Statistique, Paris.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21, 53-61.

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

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

Matei, A., and Tillé, Y. (2006). The R 'sampling' package. *European Conference on Quality in Survey Statistics*, Cardiff.

Rousseau, S., and Tardieu, F. (2004). *La macro SAS CUBE d'échantillonnage équilibré - Documentation de l'utilisateur*. Technical report, INSEE, France.