Optimal Allocation in the Multi-way Stratification Design for Business Surveys

Paolo Righi, Piero Demetrio Falorsi[1]

Abstract: Commonly, the business surveys produce estimates for a huge number of domains that define two or more partitions of the target population. When domain indicator variables are known at population level then a multi-way (or incomplete) stratification design can be used, guaranteeing a sample with planned size in each domain. The multi-way approach has some advantages with respect to the standard approach (using a one-way stratified design where the strata are obtained combining the domains of the partitions) such as: the sample allocation is more efficient (smaller sample size with same sampling errors); the response burden is reduced both in a given survey occasion and considering several survey occasions (for the combining strata with small population sizes the one-way design selects with high probability or sometime with certainty some business units in each survey occasion producing a great statistical burden). The paper shows an algorithm for defining an optimal sample allocation for the multi-way design according to the definition proposed by Bethel (1989). The procedure is suitable in the multivariate-multidomain case and assumes that the multi-way random sample selection is performed by the cube algorithm (Deville and Tillé, 2004).

1. Introduction

Commonly, the business surveys produce estimates for a huge number of domains. These domains generally define not nested partitions of the target population. When the domain indicator variables are available for each sampling unit at the sampling framework level, there are some advantages to plan a sample covering each domain. A standard approach is to use a stratified random sampling design in which strata are identified by the cross-classification of variables defining the different partitions. Multi-way (or incomplete) stratified sampling design is a second approach. This design takes under control the sample size in all the domains without using cross-classified strata.
The multi-way approach has some advantages with respect to the standard approach such as: the sample allocation is more efficient (smaller sample size with same sampling errors); the response burden is reduced both in a given survey occasion and considering several survey occasions. The problem is well known in the Structural Business Surveys where many cross-classified strata have small population size. In such cases the sampling design gives high probability or sometime certainty to some business units to be selected in each survey occasion producing a great statistical burden. Multi-way stratified sampling design avoids this problem. There are several methods implementing multi-way design, but, usually, in large scale surveys they have problem of application (Falorsi *et al.*, 2006). This is not the case of the *cube method* (Deville and Tillé, 2004) that may select a random sample of multi-way stratified design for a large population and a lot of domains. The aim of the paper is to delineate a procedure defining the set of inclusion probabilities such that the overall sample size is minimized guaranteeing that the sampling variances are lower than prefixed level of precision thresholds following the definition of optimal sample allocation given by Bethel (1989). The procedure is based on two phases: the *optimization phase* implementing an original optimal allocation algorithm, the *calibration phase* using a calibration procedure to obtain the final inclusion probabilities such that summing up on each domain an integer is achieved. The paper is devoted to the optimization phase.

2. Sampling strategy

We denote by U the reference population of N elements and by U_d (d=1, …, D) the domains where the parameters $t_{(dr)} = \sum_{k \in U} y_{rk} \gamma_{rk}$ have to be estimated, being y_{rk} the value of the r-th variable of interest in the k-th population unit and γ_{dk} the U_d domain membership indicator variable value, being $\gamma_{dk} = 1$ if $k \in U_d$ and $\gamma_{dk} = 0$ otherwise. The aim of the sampling strategy is of basing each estimate on a planned sample size. We consider a general random sampling design where are defined the minimal planned subpopulations U_h (h=1, …, H) of size N_h. That means no subpopulations completely included in U_h have a sample size fixed in the sampling selection. We assume two cases $U_d = U_h$ or $U_d = \cup_{h \in \Gamma_d} U_h$ where Γ_d is a subset of $\{1,...,H\}$. We indicate by δ_{hk} the U_h domain membership indicator variable value, being $\delta_{hk} = 1$ if $k \in U_h$ and $\delta_{hk} = 0$,

[1] ISTAT, Via C. Balbo 16, 00184 Roma, parighi@istat.it, falorsi@istat.it .

otherwise. Given the vector $\boldsymbol{\delta}'_k = (\delta_{1k},...,\delta_{Hk})$ one or more elements may assume value one. The parameters of interest are estimated by the Horvitz-Thompson estimator $\hat{t}_{(dr)}$. With this quite general sampling strategy Deville and Tillé (2005) proposed an approximated expression of the variance

$$\dot{V}_p(\hat{t}_{(dr)}\,|\boldsymbol{\pi}) = f\left[\sum_{k\in U}(1/\pi_k - 1)\eta^2_{(dr)k}\right] \tag{1}$$

where $\boldsymbol{\pi}' = (\pi_1, ...\pi_k,..., \pi_N)$ is the inclusion probabilities vector, $f = N/(N-H)$, $\eta_{(dr)k} = y_{rk}\,\gamma_{dk} - \pi_k g_{(dr)k}$ and $g_{(dr)k} = \boldsymbol{\delta}'_k\,\mathbf{B}_{(dr)}$, being

$$\mathbf{B}_{(dr)} = [\sum_{k\in U}\pi_k^2\boldsymbol{\delta}_k\boldsymbol{\delta}'_k(1/\pi_k - 1)]^{-1}\ \sum_{k\in U}\pi_k\gamma_{dk}\boldsymbol{\delta}_k y_{rk}\ (1/\pi_k - 1)\,. \tag{2}$$

In case of simple random sampling design with $U_h = U$, $H=1$ and $\pi_k = n/N$ the (1) is the exact expression of the variance. When the $\boldsymbol{\delta}'_k = (0,...,1,...,0)$ type vectors are used ($U_d = U_h$) and $\pi_k = n_h / N_h$ a stratified design is implemented. The (1) may defined as

$$\dot{V}_p(\hat{t}_{(dr)}\,|\boldsymbol{\pi}) = f\left[\sum_{h\in\Gamma_d}\frac{N_h - n_h}{n_h}\sum_{k\in U_h}\left(y_{rk} - \frac{t_{(hr)}}{N_h}\right)^2\right] = f\left[\sum_{h\in\Gamma_d}\frac{N_h - n_h}{n_h}S^2_{(hr)}\right]\,. \tag{3}$$

The (3) approximates the stratum variances according to the approximation $f(1/N_h) \approx 1/(N_h - 1)$.
Finally, when the $\boldsymbol{\delta}_k$ vectors have more the one element equal to 1 we have a multi-way stratified design.

Example: The estimates have to be obtained for three domain types T_l (l=1, .., 3). Each domain type defines a partition of the population of D_l cardinality being $D = D_1 + D_2 + D_3$. Different sampling design allows to plan the sample size of the interest domain:

- the standard approach define the U_h's combining the population of the three domain types. Then $H = D_1 \times D_2 \times D_3$ and the $\boldsymbol{\delta}'_k$ are defined as (0,..,1,...,0) vectors. We denote these design as cross-classified or one-way stratified design;
- the U_h's are defined combining all the couples of domain types. Then $H = (D_1 \times D_2) + (D_1 \times D_3) + (D_2 \times D_3)$;
- some U_h's agree with the domains of one population partitions (for instance T_1) and the others U_h's are defined combining couples of the remaining domain types (T_2 and T_3). Then $H = D_1 + (D_2 \times D_3)$;
- the U_h's agree with the domains of populations. Then $H = D_1 + D_2 + D_3$.

The choice of the sampling design depends on theoretical and operative reasons. In particular, from the operative view point to implement random selection scheme to obtain the one-way stratified design is quite straightforward, while is more difficult to define a random selection for the multi-way stratified design. In 2004 Deville and Tillé proposed the cube algorithm suitable to select randomly a multi-way stratified sample.

3. Optimal allocation algorithm

The sampling design needs to know the elements of the $\boldsymbol{\pi}$ vector. We propose an algorithm for the definition of an optimal inclusion probability vector $\boldsymbol{\pi}^*$ according to the following optimality criterion:

$$Min(\sum_{k\in U}\pi^*_k) \text{ such that: (a) } \dot{V}_p(\hat{t}_{(dr)}\,|\boldsymbol{\pi}^*) \le \overline{V}_{(dr)} \ \ \forall\ (dr);\ \text{(b) } 0 < \pi^*_k \le 1, \tag{4}$$

being $\overline{V}_{(dr)}$ a fixed variance threshold for the domain U_d on the *r*-th variable of interest. The search of the π_k^*, denoted by *optimization phase*; is coupled to a *calibration phase*. The last phase changes as little as possible the optimal probabilities in the calibrated probabilities in such a way that summing up on each domain the calibrated probabilities gives an integer. In the paper the optimization phase is described. For calibration phase see Falorsi and Righi (2008). The optimization phase implements an algorithm for solving the non linear programming problem (4). Nevertheless, we note the (4a) constraints depend on the unknown variables of interest. In practice only model predicted values can be used. We show the algorithm in this operative context. We consider a general prediction model *M*

$$\begin{cases} y_{rk} = \tilde{y}_{rk} + u_{rk} \\ E_M(u_{rk}) = 0 \forall k; E_M(u_{rk}^2) = \sigma_{rk}^2; E_M(u_{rk}, u_{rl}) = 0 \ \forall k \neq l \end{cases}.$$

We assume σ_{rk}^2 as known. To take into account the model uncertainty, the (1) is replaced by the Anticipated Variances (Isaki and Fuller, 1982) and the constraints (4a) are defined as $AV_p(\hat{t}_{(dr)} \mid \boldsymbol{\pi}^*) \leq \overline{V}_{(dr)}$. An upward approximation of the anticipated variances for the proposed strategy is

$$AV_p(\hat{t}_{(dr)} \mid \boldsymbol{\pi}^*) \approx f \sum_{k \in U} (1/\pi_k^* - 1) \tilde{\eta}_{(dr)k}^2 + \sum_{k \in U} (1/\pi_k^* - 1) \gamma_{(dr)k} \sigma_{rk}^2 , \qquad (5)$$

where $\tilde{\eta}_{(dr)k}^2$ is computed by means of a model predicted value $\tilde{y}_{rk}$. The approximation neglects a residual term that we do not show for sake of brevity. However, the optimization procedure does not change if the corrected anticipate variance is taken into account. To obtain a solution of the optimization problem we formulate constraints (5) as

$$\sum_{k \in U} \gamma_{dk} (\tilde{y}_{rk}^2 + \sigma_{rk}^2) / \pi_k^* \leq \overline{V}_{(dr)} + \sum_{k \in U} \gamma_{dk} (\tilde{y}_{rk}^2 + \sigma_{rk}^2) + C_{(dr)}(\boldsymbol{\pi}^*, \tilde{\mathbf{g}}_d)$$

being $C_{(dr)} = f \left[\sum_{k \in U} 2(1 - \pi_k^*) \gamma_{dk} \tilde{y}_{rk} \tilde{g}_{dk} - \sum_{k \in U} \pi_k^* (1 - \pi_k^*) \tilde{g}_{dk}^2 \right]$, where $\tilde{\mathbf{g}}_d' = (\tilde{g}_{1k}, ..., \tilde{g}_{dk}, ..., \tilde{g}_{Dk})$, $\tilde{g}_{dk} = \boldsymbol{\delta}_k' \tilde{\mathbf{B}}_{(dr)}$ with $\tilde{\mathbf{B}}_{(dr)}$ given by (2) replacing y_{rk} with $\tilde{y}_{rk}$. The optimization phase is performed by the proposed algorithm:

1. Initialization: for $\alpha = 0$, let ${}^{(\alpha=0)}\pi_k = 1/\overline{n}$ (k=1,…,N) be the initial values of the inclusion probabilities, in which $2D \leq \overline{n} < N$ is a fixed quantity; $\overline{n} = 2D$ is a reasonable choice;
2. Iteration over α. for α = 0,1,2,3,…, calculate ${}^{(\alpha)}\tilde{g}_{dk}$;
3. Iteration over τ:
   a. *Initialization*: for τ =0 let ${}^{(\alpha,\tau=0)}\pi_k = {}^{(\alpha)}\pi_k$;
   b. *Calculation*: for τ =0,1,2,3, calculate
   ${}^{(\alpha,\tau)}C_{(dr)} = f \left[\sum_{k \in U} 2(1 - {}^{(\alpha,\tau)}\pi_k) \gamma_{dk} \tilde{y}_{rk} \, {}^{(\alpha)}\tilde{g}_{dk} + \right.$
   $\left. - \sum_{k \in U} {}^{(\alpha,\tau)}\pi_k (1 - {}^{(\alpha,\tau)}\pi_k) ({}^{(\alpha)}\tilde{g}_{dk}^2 \right]$;
   c. *Updating*: calculate ${}^{(a,\tau+1)}\pi_k$ such that:
   $\sum_U {}^{(\alpha,\tau+1)}\pi_k$ is minimized with
   $\sum_{k \in U} \gamma_{dk} (\tilde{y}_{rk}^2 + \sigma_{rk}^2) / {}^{(\alpha,\tau+1)}\pi_k \leq$
   $\overline{V}_{(dr)} + \sum_{k \in U} \gamma_{dk} (\tilde{y}_{rk}^2 + \sigma_{rk}^2) + {}^{(\alpha,\tau)}C_{(dr)}({}^{(\alpha,\tau)}\boldsymbol{\pi}, {}^{(\alpha)}\tilde{\mathbf{g}}_d)$
   $0 < {}^{(\alpha,\tau+1)}\pi_k \leq 1$ (K =1, …, N). The optimization is performed by slight modification of the algorithm proposed by Chromy (1987). Technical details in Falorsi and Righi (2008);

d. *Exit rule*: let ε denote a fixed small quantity. If

$$\sum_U \left| {}^{(\alpha,\tau+1)}\pi_k - {}^{(\alpha,\tau)}\pi_k \right| \leq \varepsilon , \tag{6}$$

the iterations over τ are stopped and the updated inclusion probabilities ${}^{(\alpha+1)}\pi_k = {}^{(\alpha,\tau+1)}\pi_k$ (k=1,…, N) are calculated. Otherwise, the *Calculation* and *Updating* steps are iterated with $\tau = \tau + 1$ until the condition (6) is satisfied.

4. General exit rule: if

$$\sum_U \left| {}^{(\alpha+1)}\pi_k - {}^{(\alpha)}\pi_k \right| \leq \varepsilon . \tag{7}$$

The algorithm ends and final inclusion probabilities are given by $\pi_k^* = {}^{(\alpha+1)}\pi_k$ (k=1,…, N). Otherwise, the Iteration over α and the Iteration over τ are iterated with $\alpha = \alpha + 1$, until the condition (7) is respected.

Remark 1. In case of one-way stratified design and when $\tilde{y}_{rk}\gamma_{dk} = \boldsymbol{\delta}_k' \tilde{\mathbf{B}}_{(dr)}$ the anticipated variance is given by

$$AV(\hat{t}_{(dr)} \mid \boldsymbol{\pi}^*) = \sum_{h \in \Gamma_d} \sigma_{rh}^2 \sum_{k \in U_h} (1/\pi_k^* - 1) , \tag{8}$$

where disappear the terms $\tilde{y}_{rk}$. In this case the original Chromy algorithm is it suitable to achieve the solution in one iteration. The optimal inclusion probabilities are such that $\pi_k^* = \pi_h$.

Remark 2. For one-way stratified design with constant inclusion probabilities in each U_h the (8) assumes the same form of the (3) except for the variance terms. In (3) these terms represent the design variances while in the (8) are the model variances. In practice the (design or model) variance terms have to be estimated and most likely these estimates will be based on the same procedure leading to the same sampling allocation.

4. Empirical evaluations

Currently, the proposed algorithm has been applied only in experimental contexts. A simulative study on real data of the *Italian Graduates' Career Survey* conducted by Italian National Statistical Institute has tested if the algorithm converges to an optimal solution. More detailed results are given in Righi and Falorsi (2011). A briefly description of the simulation is the following: the survey produces estimates for 8 types of domains with two very detailed not nested domain partitions: degree by sex (first partition) and university by subject area degree (second partition). Using the cross-classified stratification design more than 7,700 strata are obtained for planning 160 plus 90 domains. A Monte Carlo simulation has been performed on a sub-population of 3,427 units covering 20 and 15 domains belonging respectively to the first and second partition. Two cases have been studied: the first assumed the interest variables as known; the second assumed they were predicted. In each case the algorithm has converged with a number of iterations depending on the exit rule constraints. We obtain 6 and 7 iteration respectively for the first and second case when we set $\varepsilon = 0$. But fixing $\varepsilon = 0.05$ three iterations over α would have been necessary to satisfy the convergence criterion in both cases. The solutions have been compared with cross-classified stratification design having 91 strata. This design shows to be inefficient with respect to the multi-way design. Adopting the same variance thresholds the one-way design increases the sample size of about 7% (and more than 10% in some domains). The bad performance depends on the too detailed stratification coupled with the constraints to have 2 units per stratum for computing unbiased variance estimates.

A second experiment has been performed on the 1999 population of the enterprises from 1 to 99 employees belonging to the Computer and related economic activities (2-digits of the NACE rev.1 classification). The data base used for the simulation study has N=10,392 enterprises. Two domain partitions have been considered: the geographical region with 20 domains; the Economic Activity Group by Size class with 24 domains. In the experiment a sample size of 360 units has been fixed (for further details see Falorsi and Righi 2008). The size is equal to the combining strata of the one-way stratified design. Then the experiment highlights the cases in which the standard design approach for planning each domain is unsuitable for budget constraints. An easy strategy is to drop one or more stratifying variables or to group some of the categories. Nevertheless, some planned domains become unplanned and some of them can have small or null sample size.

The aims of the Monte Carlo simulation have been: verify the performances of the random selection scheme with a large population and many domains; compare some sampling strategies based on a multi-way stratified design, feasible in a large scale surveys and the one-way stratified designs where strata are given alternatively by the domains of a singular

partition. A compromise (not optimal) sample allocation has been used and the calibration phase has been performed. The simulation underlined the selection based on the cube algorithm is feasible and that controlling the sample on the small domains by means of multi-way design gives a significant improvement on the accuracy of the estimates.

5. Conclusion

Multi-way (or incomplete) stratification design is a useful sampling approach when the planning of the sample size for domains belonging to different partitions of the population is required. Usually, to achieve this task, the standard approach is based on a stratification given by the combination of these domains (cross-classified or one-way stratification) because the random selection scheme is straightforward to implement. Nevertheless, the approach may have some drawbacks: the stratification has not the aim to improve the efficiency of the estimates; the sample allocation may not be optimal due to minimum sample size constraints in each stratum; statistical burden may occur in small strata; too detailed stratification could need a too large sample with respect to the budget constraints.

These problems arise especially in the business survey in the Official Statistics. In fact, the European *Council Regulation* on Structural Business Statistics establishes that the parameters of interest refer to estimation domains defined by three different partition of the enterprise population. In the Italy case about 1,800 estimation domains are defined while the number of non-empty strata of the cross-classification design is larger than 37,000. Therefore, the design require a large sample only to cover each stratum.
On the other hand, the multi-way stratification design often has as main trouble the implementation of a random selection scheme especially with large population. Recently, the cube algorithm, proposed Deville and Tillé (2004), overcomes such drawback and the design may be applied to the large scale surveys as well.

The paper focuses the sampling allocation issue in the multi-way design. An algorithm for computing the optimal inclusion probabilities, which is the optimal sample allocation, is defined. The algorithm implements the allocation for a general multi-way sampling design in which the standard approach (one-way stratification) is a special case. Moreover, the allocation is multi-domain and multivariate: the sample size is minimized guaranteeing that the sampling variances of the target estimates of several variables on the planned domains are lower than prefixed level of precision thresholds. As for other common allocation procedures, the proposed method requires the knowledge of the variables of interest while in practice only predictions may be available. Then the algorithm takes into account the prediction uncertainty.

Some experiments, shortly described in the paper, have been performed. They have several tasks: verify the algorithm in terms of convergence; compare the sample size of the multi-way stratification with respect to a one-way stratification design; confirm the multi-way random selection algorithm is suitable when many planned domains are involved; evaluate the improvement of accuracy of the small domains when the sample sizes are planned by means of the proposed method. The simulations have given satisfactory results.

Further experiments are need especially for stressing the allocation procedure on large data sets. In particular, to have some indication on the computational effort and to confirm the proposed algorithm converges to an optimal solution.

References

Bethel J. (1989) Sample Allocation in Multivariate Surveys, *Survey Methodology*, 15, 47-57.

Chromy J. (1987). Design Optimization with Multiple Objectives, *Proceedings of the Survey Research Methods Section. American Statistical Association*, 194-199.

Deville J.-C., Tillé Y. (2004) Efficient Balanced Sampling: the Cube Method, *Biometrika*, 91, 893-912.

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

Falorsi P. D., Righi P. (2008) A Balanced Sampling Approach for Multi-way Stratification Designs for Small Area Estimation, *Survey Methodology*, 34, 223-234.

Falorsi P. D., Orsini D., Righi P., (2006) Balanced and Coordinated Sampling Designs for Small Domain Estimation, *Statistics in Transition*, 7, 1173-1198.

Isaki C.T., Fuller W.A. (1982) Survey design under a regression superpopulation model, *Journal of the American Statistical Association*, 77, 89-96.

Righi P., Falorsi P. D., (2011) Optimal Allocation Algorithm for a Multi-Way Stratification Design, *Proceedings of the Second ITACOSM Conference*, 27-29 June 2011, Pisa, 49-52.