Optimal sample allocation for the Incomplete Stratified Sampling design

Claudia De Vitiis[1], Paolo Righi[2], Marco Dionisio Terribili[3]

Sommario

Per indagini campionarie finalizzate alla produzione di stime per una molteplicità di domini di stima, un disegno di campionamento ampiamente adottato è il disegno stratificato in cui gli strati sono definiti sulla base dell'incrocio delle variabili che definiscono i domini di stima. Nei casi in cui le variabili di stratificazioni sono non annidate e presentano molte modalità, il disegno stratificato può risultare inefficiente. Nel lavoro si prospetta la definizione di un disegno campionario detto a Stratificazione Incompleta che sfrutta tutto il potenziale informativo ausiliario disponibile sia dalla lista di campionamento sia da altre fonti quali indagini precedenti: in tal modo la dimensione campionaria è fissata per i domini di interesse ottenendo le precisioni desiderate per le stime campionarie, conseguendo una riduzione della dimensione campionaria complessiva a parità di precisioni attese, poiché il processo di allocazione non ha vincoli di numerosità negli strati.

Parole chiave: stratificazione a più vie, allocazione campionaria

Abstract

For sampling surveys aiming at producing estimates for different domains of interest, a sampling design widely adopted in official statistics is the Stratified Simple Random Sampling (SSRS) design in which strata are defined by crossing of the variables that define the domains of estimate. When there are many strata, the SSRS design could be inefficient. We propose an alternative sampling design denoted as Incomplete Stratified Sampling (ISS) design. The design exploits all the potential auxiliary information available both from the sampling frame and from other sources such as previous surveys in a more efficient way with respect to the traditional SSRS design. The ISS design enable to fix the sample size for the each estimation domain obtaining the required precisions for the sample estimates, achieving a reduction of the overall sample size with respect to the SSRS design since for the latter the allocation process has no constraints on stratum sample sizes.

Keywords: multi-way stratification, sample allocation

1. Introduction

Literature on finite population sampling has devoted much attention on planning the sampling design and the outlining of the inclusion probabilities. The paper takes into account the class of stratified designs and in particular the Stratified Simple Random Sampling (SSRS) designs. In SSRS designs the definition of the inclusion probabilities coincides with the sample allocation by stratum, being the number of stratum sampled units given by summing up the inclusion probabilities over the stratum population. These designs are broadly applied in the official statistics: firstly for the easy implementation, secondly because they can be used to plan the sample size of sub-populations or domains of interest at design stage allowing to control the sampling errors in this phase. For the latter purpose, the domains of interest are classified by type of domain. For instance, in the socio-demographic surveys the partition types could be the gender, the province or region of residence, the age by class. Such partitions could be nested (for instance province in the region) or not nested (for instance gender and age by class). A practical SSRS design considers the finer not nested partitions and combines the category of each partition for obtaining the strata. In this way, the sample size of each domain is planned because are planned the stratum sample sizes. These designs are sometimes denoted as multi-way stratified de-

[1] Ricercatore Istat, e-mail: devitiis@istat.it
[2] Ricercatore Istat, e-mail: parighi@istat.it
[3] Collaboratore Tecnico di Ricerca Istat, e-mail: terribili@istat.it

sign (Winkler, 2009) and, in particular, if the stratification is built up by two partitions we have a 2-way stratification design. Usually and mainly the instrumental role (plan the domain sample sizes) of the multi-way strata outweighs the efficiency issues of a sampling design.

The allocation of a SSRS design can be implemented according to an optimization problem. The optimal allocation for a univariate population is well-known (Cochran, 1977). In case of a multivariate scenario, where more than one characteristic is to be measured on each sampled unit, the optimal allocation for individual characteristics do not have much practical use, unless the characteristics under study are highly correlated. This is because an allocation that is optimal for one characteristic will generally be far from optimal for others. Therefore, the criteria established for the problem's multidimensionality leads to a definition of an allocation that loses precision, compared to the individual optimal allocation. For these reasons, the methods are sometimes referred as compromise allocation methods (Khan et al., 2010). Although we do not talk about optimal allocation we still define reasonable sample allocation criteria. They depend on several elements defining the sampling strategy: the inferential approach, the parameters of interest, the domains of interest, the estimator and, finally, the a priori information on the phenomena of interest. To tackle the problem several compromise allocation criteria have been proposed. A classical compromise allocation is given by the convex function of proportional allocation to population sample size and equal stratum sample size allocation (Costa et. al. 2004) or the power allocation (Bankier, 1988). Chromy (1987), Bethel (1989) and Choudhry et al. (2012) give a mathematical formalization to the compromise allocation, according to an optimization problem. All these criteria are suitable for the SSRS design. Along with the SSRS design, in this paper we propose another sampling design that we denote as *incomplete stratification sampling* (ISS) design (Righi and Falorsi, 2008; 2011). The ISS design is based on a stratification, where the units belongs to the same stratum have the same inclusion probabilities, but, differently from the SSRS design, the number of sampled units is a random variable while the interest domain sample sizes are still planned at design stage. The ISS can be considered a special case of balanced sample in the randomization approach (Deville and Tillé, 2004), where the balancing totals are the resulting domain allocations. are This feature could have a strong impact on the overall sample dimension. On the other hand the sample allocation for the SSRS design requires at least two sampled units in each stratum (if two in the population) to obtain unbiased variance estimates and the inclusion probabilities in each stratum must be rounded off such that summing up at stratum level we obtain an integer number (so that we can select an integer number of sampled units). These two issues are not strictly related to the optimization problem defining the compromise allocation and they represent a sort of exogenous constraints that produces inefficiency on the allocation. These problems can be overcome by the ISS design.

In section 2 we give a brief formalization of the optimization problems for the SSRS and ISS sampling design in the multivariate scenario. We show that the two formalizations are quite similar. Section 3 focuses on the definition of some input parameters involved in the optimization problem. They can significantly modify the optimal sample allocation solution. We compare the allocations achieved by a SSRS and ISS designs in section 4 where an experiment on *University graduates' vocational integration* survey data is performed. Some conclusions are presented in section 5.

2. Allocation problem

Let U be the reference population of N elements and let U_d (d=1, …,D) be an estimation domain, i.e. a generic sub-population of U with N_d elements, for which separate estimates must be calculated. Furthermore we denote by U_h the h^{th} (h=1, …, H) sub-population where the inclusion probability π_k of unit k (k=1, …, N) must be equal to π_h. In the SSRS design U_h is a stratum and each U_h does not cut across the U_d's. The allocation problem searches for the vector $\boldsymbol{\pi}' = (\pi_1,...,\pi_h,...\pi_H)$ satisfying a given criterion.

We formalize the criterion according to an optimization problem. Both for the SSRS and ISS designs it is mainly based on the following system

$$\begin{cases} Min\left(\sum_{k\in U} \pi_h\, c_h\right) \\ V(\hat{t}_{(dr)}) \le \overline{V}_{(dr)} \quad (d = 1,...,D; r = 1,...,R) \\ 0 < \pi_h \le 1 \qquad\qquad (h = 1,...,H) \end{cases} \quad (2.1)$$

where: c_h is the uniform cost for collecting information from unit $k \in U_h$;

$V(\hat{t}_{(dr)})$ is a measure of precision (variance) of the estimate $\hat{t}_{(dr)}$ of total $t_{(dr)} = \sum_{k \in U_d} y_{kr}$ on the domain U_d for the variable y_r, in which the expression of $V(\hat{t}_{(dr)})$ depends on the sampling design implemented; $\overline{V}_{(dr)}$ is a fixed precision threshold for $\hat{t}_{(dr)}$ estimate; the y_r ($r = 1, \ldots, R$) are the driving variables for the allocation. In this formalization their totals represent the (main) parameters of interest.

In case of the SSRS design further constraints are necessary:

$$\begin{cases} \boldsymbol{\pi}_h = 1 \text{ when } N_h = 1 \\ N_h \boldsymbol{\pi}_h \geq 2 \text{ when } N_h \geq 2 \\ N_h \boldsymbol{\pi}_h \text{ must be equal to an integer} \end{cases} \qquad (2.2)$$

If the ISS design is considered we have the following constraints

$$\sum_{U_h \in U_d} N_h \boldsymbol{\pi}_h \text{ must be equal to an integer} \qquad (2.3)$$

The optimization problem (2.1) with the constraints (2.2) or (2.3) plans the U_d sample sizes so that is minimized the expected cost ensuring that the precision measures on the estimates of the driving variables are bounded and that the inclusion probabilities lie between 0 and 1.

For a concrete use of the optimization problem other parameters, included in the $V(\hat{t}_{(dr)})$ expression, have to be fixed. In particular: the definition of $V(\hat{t}_{(dr)})$, in the SSRS design, requires the knowledge of the variance S_{hr}^2 for the variable y_r in the stratum U_h; in the ISS design the population mean $\overline{Y}_{hr}$ for each variable y_r in the stratum U_h has to be known as well. Of course such parameters are unknown as they are the targets of the survey. Then, we have to replace these values with some estimates and to treat the estimates as true values. A common strategy is to use the previous survey data where the variable y_r have been collected and to perform an estimation procedure.

The estimation of S_{hr}^2 and $\overline{Y}_{hr}$ is crucial on the final allocation and at the same time often underrate when planning the sampling design.

Chromy (1987), Bethel (1989), Falorsi *et al*. (1998) and Choudhry *et al*. (2012) propose different algorithm converging to the same solution for solving the problem (2.1) when $V(\hat{t}_{(dr)})$ is the variance of the SSRS design. Righi and Falorsi (2008; 2011) consider the variance expression of the ISS design in the optimization problem and propose a new algorithm. Since the ISS is a special case of the balanced sampling design, where the balancing variables are ${}_{\mathrm{d}}\delta_{\mathrm{k}}\pi_{\mathrm{k}}$ (being ${}_{\mathrm{d}}\delta_{\mathrm{k}}$ the variable indicator of domain *d*), the expression for the variance proper for the balanced sampling (Deville and Tillé, 2005) is taken into account in the allocation procedure the allocation procedure.

3. Estimation of the parameters for the allocation

The section focuses on the estimates of the $\overline{Y}_{hr}$ and S_{hr}^2 for the allocation. We assume that the U_h are small domains and direct estimates based on previous survey data are not reliable. For this reason, the practical approach is to use a model based approach borrowing strength from larger sub-population data. The aim is to exploit as much as possible the knowledge on the y_r variables before conducting the survey, because in this way a sample size as small as possible will be enough for obtaining satisfying estimates of such characteristics. We consider $\overline{Y}_{hr}$ as a model prediction of each value y_{kr} for $k \in U_h$, being the auxiliary variables of the model known also in the list frame available for the sampling selection; S_{hr}^2 are the model variance. Therefore, the first step for setting up the optimization problem is to produce the *best* prediction of $\widetilde{\overline{Y}}_{hr}$ and $\widetilde{S}_{hr}^2$. What *best* means is strictly related to the goodness of fit of the estimated

model with the previous survey data. According to this approach we can go beyond the multi-way stratification. In fact, the best prediction model for the y_{kr} could be defined out of the multi-way strata so that the mean and variance model can be different within the multi-way strata. In this sense we are searching the optimal stratification (Khan *et al.*, 2008) with the only constraints that the strata do not cut across the domains of estimate for guaranteeing that the domain sample sizes are planned at the design stage. Furthermore, we could have an individual prediction value when using a prediction model with at least one continuous auxiliary variable.

We point out that the granularity of the stratification affects the final allocation, especially when a SSRS is adopted, since the weight of the constraints (2.2) increases in the optimization problem when the number of small strata increases.

In the following, an application on real survey data tests the sample allocation issue with the SSRS and ISS under different prediction models leading to the multi-way stratification or a more detailed stratification. We restrict the analysis to fixed effect models but in general random effect models typically used for the small area estimation problem could be investigated (Rao, 2003).

The output of the optimization procedure gives a sample allocation with the expected percentage CV for the estimates on the domains. These values will be lower or equal to the CV thresholds. In practice, when the sampling survey has been conducted and the estimates computed, the real CV estimates (in absence of non-response) will generally differ from the expected ones for two main reasons: the super-population models generating the variable of interest differ from the models used for defining the input parameters; the input parameters are estimated, rather than being true. When we search for a best model, we try to choose a model as closest as possible to the true super-population model. In this way, we can reduce the possible difference among the expected and the observed CV of the estimates.

4. Application

The experiment has been carried out on the basis of data from the last edition of the university graduates' vocational integration survey conducted by the Italian National Statistical Institute.

The survey aims at investigating the graduates' employment conditions, the working stability, the job placement and the economic activity area. The data have been collected in 2011 on the population of about 173,800 graduates', who hold a Bachelor's Degree during the calendar year 2007. The next planned edition of the survey will be conducted during year 2015 and it will regard the population graduates', who got a university degree, both Bachelor and Master, during the calendar year 2011.

The interest domains of the survey are defined on the basis of gender, degree programs and university, variously crossed and aggregated. The 2011 survey used a SSRS design where the 2,981 not empty strata were obtained by crossing the variables degree program, gender and university.

The application has been carried out on 2011 survey data in order to plan the sample design of 2015 survey edition. Two types of domains are considered: degree programs crossed with gender (DOM1) and university crossed with educational area (DOM2), for an overall number of 542 domains. The survey produces actually estimates for other more aggregate domain partitions, which can be obtained as aggregation of DOM1 and/or DOM2.

The experiment has been developed in two main phases: the first one devoted to the selection models for predicting the $\bar{Y}_{hr}$ and S^2_{hr}, based on 2011 survey and frame data. In the second phase the SSRS and ISS allocations have been compared in terms of overall sample sizes.

The first phase used 2011 complete information, deriving from both survey and frame, to estimate model parameters, to be used for planning the next edition of sample design for which only auxiliary information in the frame is available.

4.1 Model selection

We consider three binary variables y_r **(*r=1,2,3*)** describing the condition of the graduates three years later than the graduation: working (yes/no), looking for a job (yes/no), studying (yes/no). To predict the binary responses, logistic regression models have been fitted using auxiliary variables chosen from the list of variables available in the previous survey and in the current sampling frame: UNIVERSITY *of the degree achievement* (80 modalities), *educational* AREA *of the course* (9 modalities), *branch of knowledge of the course* or GROUP (16 modalities), *degree program* or COURSE, AGE CLASS *at the graduation moment* (3 modalities), *NUTS 2 residence* REGION (21 modalities), GENDER (2 modalities) and FINAL GRADES CLASS (3 modalities). The original continuous variables, age and final grades, have been recoded as categorical variables to allow the implementation of both SSRS and ISS designs.

Several logistic regression models have been studied relatively to the three dependent variables and the *Akaike Information Criterion* (AIC) has been used to evaluate their goodness of fit (table 4.1). The investigated models have different and increasing levels of complexity. Models from 1 to 3 are the simplest ones and they are considered as a

benchmark for the more complex ones. Model 4 was the one used for planning the 2011 sample design, the previous survey occasion.

Model 5 uses all the auxiliary variables defining the planned domains (gender, university, educational area) but aggregating them, in order to deal with computational issues.

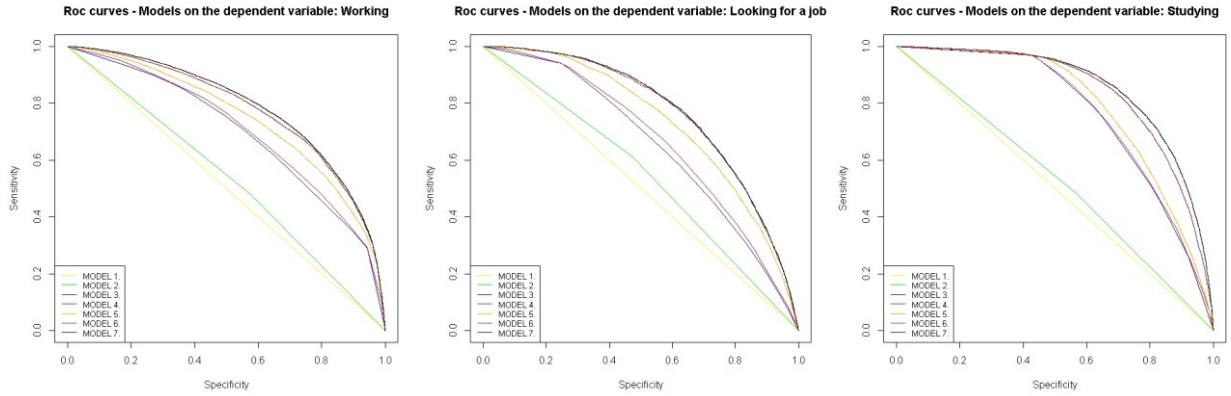
Models 6 and 7 have been chosen according to the goodness of fit; they differ for the variable GROUP (model 6) and the COURSE (model 7). These models have been studied with the aim to describe accurately the dependent variables and the obtained predictions vary within the two-way strata. In these two models, the units with the same covariate pattern (or profile) have the same prediction. In the allocation procedure each profile is a stratum.

Table 4.1 – Proposed models' AIC, relatively to the dependent variable working, looking for a job, studying.

Model	AIC-Working	AIC-Looking for a job	AIC-Studying
1: Total average (only intercept)	37,700	26,570	42,976
2: Gender	37,624	26,451	42,892
3: Group	34,251	25,256	33,885
4: Gender+ Group+ Group * Gender	34,020	25,088	33,782
5: Gender+Area + Gender*Area + University	32,737	23,865	32,941
6: University+Group+Age class+Region+Gender+Final grades class	30,390	22,252	29,531
7: University+Course+Age class+Region+Gender+Final grades class	30,004	22,231	28,577

Table 4.1 shows that the increasing complexity produces decreasing AIC, denoting a better fit. The relative differences among the model goodness of fit are depicted by the ROC curves (figure 4.1).

Figure 4.1 – Proposed models' ROC curves, relatively to the dependent variable working, looking for a job, studying.



In the graphs sensitivity (true positive rate) is plotted in function of specificity (false positive rate), varying the cut-off point; so each point on every ROC curve represents a sensitivity/specificity pair. The area under the ROC curve (called AUC, in acronym form) is a measure of how well a model can distinguish between two modalities of a dependent variable (working/not working, looking/not looking for a job, studying/not studying). The more complex the model is, the more bent the curve is, maximizing its AUC.

The graphs confirms that the model 7 is the best model, and so it can be considered the closest one to the true and unknown superpopulation model generating the three variables of interest.

4.2 Sample allocation

Once we got the predicted values of the needed quantities discussed in section 3 through the models described in section 4.1, we compared the sample allocation of the optimization problem (2.1) using the constraints (2.2) or (2.3) respectively for the SSRS and ISS design. Both the sample allocations were performed fixing the same precision thresholds according to the percentage Coefficient of Variation (CV) of the sampling estimates for the totals of the three variables of interest. For DOM1 domain type the following three CV had been considered: 13%, 25% and 20% respectively for "working", "looking for a job", "studying"; for DOM2 domain type the following three CV had been considered: 13%, 25% and 15%.

4.3 Results

Table 4.2 shows the overall sample sizes for both the sampling designs, SSRS and ISS, having set the cost for collecting information constant. Furthermore, the table displays the number of strata considered in the designs. For models 1 to 5, where the profiles are aggregations of the two-way strata, we have 2,981 strata. Models 6 and 7 define respectively 8,743 and 31,486 profiles so, therefore, strata.

Tavola 4.2 - Number of strata for the proposed models and sample sizes for SSRS and ISS designs

Model	*Strata considered in the allocation procedure*	SSRS	ISS
1: Total average (only intercept)	2,981	26,419	24,845
2: Gender	2,981	26,673	25,232
3: Group	2,981	31,539	30,061
4: Gender+ Group+ Group * Gender	2,981	31,345	29,879
5: (Gender*Area)+University	2,981	36,624	35,027
6: University+ Group+Age class+Region+Gender+Final grades class	8,743	63,246	34,620
7: University+Course+Age class+Region+Gender+Final grades class	31,486	63,168	34,622

The comparison between SSRS and ISS allocation shows that the latter design requires a smaller sample size to satisfy the precision thresholds. What happens in the SSRS design is that the constraints (2.2) enlarge the sample size with the result that the expected CVs can result unnecessarily below the threshold stated than the expected CVs obtained for the ISS design.

The further interesting evidence is related to the model choice. Table 4.2 displays that the simplest model 1 gives the smallest sample size both for the SSRS and ISS design. The result does not imply that we have to choose model 1, but that the allocation for model 1 will give observed CV estimates probably very far from the expected ones.

5. Conclusions

The sampling surveys in official statistics are usually characterized by a large number of domains for which several parameters have to be estimated. When the domain membership binary variable values are known for each population unit at the design stage it could be useful to select a sample in which the sample size for each domain is planned. In this way, in some extent the design enables to control the sampling errors of the domain estimates. The paper introduces the Incomplete Stratified Sampling (ISS) design to deal with the domain sample size allocation and compares the ISS efficiency in terms of overall sample size to the efficiency of the multi-way Stratified Simple Random Sampling (SSRS) design commonly used to fix the number of domain sampled units at design stage. The comparison is carried out using optimal allocation methods that, in the case of multivariate and multi domain context, actually define a compromise allocation criterion. The methods have been evaluated modifying the mean and variance input parameters. The modifications depend on the working models used for predict these parameters since in practice they are unknown. The estimated or predicted parameters are used as if they were observed and, as a consequence, if the estimated values are too far from the true values the allocation can lead to misleading conclusion on the expected precision of the estimates. The main results of the experiments reveal that the ISS design always outperforms the SSRS especially when the number of strata increases. That means the ISS is a more flexible tool and it can be used to choice the best working model to predict the input parameters. On the other hand, when the SSRS design has to be implemented we must pay attention on the number of strata generated by the working model to avoid the sample size inflates too much because of exogenous design constraints.

Finally, the next 2015 edition of the university graduates' vocational integration survey will be realized using the ISS design and this choice allow to define a more efficient design than in the past.

References

Bethel, J. (1989). *Sample allocation in multivariate surveys.* Survey Methodology, 15, 47-57.

Bankier, M. (1988). *Power allocation: Determining sample sizes for sub-national areas.* The American Statistician, 42, 174-177.

Choudhry, G.H, Rao, J.N.K., Hidiroglou, Michael A. (2012). *On sample allocation for efficient domain estimation.* Survey Methodology, pp. 23-29.

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

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York : John Wiley & Sons, Inc..

Costa, A., Satorra, A. and Ventura, E. (2004). *Using composite estimators to improve both domain and total area estimation.* SORT, 28, 69-86.

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., Ballin M., De Vitiis C., Scepi G. (1998). *Principi e metodi del software generalizzato per la definizione del disegno di campionamento nelle indagini sulle imprese condotte dall'ISTAT*, Statistica Applicata, 10, 235-257.

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., Righi P. (2008). *Optimal Allocation in the Multi-way Stratification Design for Business Surveys*. Proceedings EESW11 2011 European Establishment Statistics Workshop. 12 –14 September 2011, Neuchâtel, Switzerland.

Khan, M. G. M., Maiti, T., Ahsan , M. J. (2010) *An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach*, Journal of Official Statistics, pp. 695–708.

Khan, M.G.M., Nand, N., Ahmad, N. (2008). *Determining the optimum strata boundary points using dynamic programming.* Survey Methodology, 34, 205-214.

Rao J. N. K. (2003). *Small Area Estimation*, Wiley, New York.