Optimal adaptive sample allocation in stratified sampling under budget constraints

Elisabetta Carfagna*
University of Bologna, Bologna, Italy – elisabetta.carfagna@unibo.it

Silvia Missiroli
Bocconi University, Milano, Italy – silvia.missiroli@phd.unibocconi.it

Several authors have faced the problem of sample allocation and selection without previous information on the variability inside the strata, suggesting various kinds of two-step sampling or sequential sampling strategies. However, proposed methods either do not allow design unbiased estimates of the population parameters or are not optimal or do not take into consideration budget constraints. In this paper, we propose a group sequential adaptive procedure with permanent random numbers that generates design unbiased minimum variance estimates of the population mean, under budget constraints. Through a Monte Carlo simulation study, we investigate the optimum combination of number of steps and sample units per step, we prove that the proposed procedure is more efficient than the ones proposed in the literature and assess the impact of various values of the cost components on the proposed procedure. We also propose an approach for identifying the optimal adaptive sequential allocation when the population distribution is unknown.

Keywords: Stratified sampling; Optimal allocation; Group sampling; Permanent random numbers; Unbiased Estimates.

1. Introduction

In survey sampling for finite populations, one of the most important challenge is to offer an efficient sampling procedure in terms of cost, time and precision, without omitting flexibility.
Adopting a good stratification with optimal strata allocations (Neyman's) allows to increase the efficiency of the estimator. However, when the strata variances are unknown, Neyman's allocation cannot be computed. In this case, adaptive sampling can be useful to gain the missing information through the results obtained along the way. Stein (1949), Chow and Robbins (1965), Ray (1957) are some of the authors who first proposed two steps or adaptive sequential procedures for infinite population. In the context of stratified finite populations, Thompson and Seber (1996) suggested an adaptive approach in K phases (phases are sampling steps) and an estimator of the population mean given by the weighted mean of the estimates at the various phases. This estimator is unbiased if the weights are fixed in advance (do not depend on observations made during the survey) and each of the strata is sampled at each phase. These two conditions have a negative impact on the efficiency of the estimator. Carfagna (2007) proposed a two steps adaptive procedure (TSPRN) with the use of permanent random numbers (Ohlsson, 1995), which pursues Neyman's allocation and allows to get unbiased and more efficient estimators than those obtained through Thompson and Seber's method when K = 2. In fact, the TSPRN procedure, at the second step, allows selecting supplementary units only in those strata where supplementary selection is necessary, not in all the strata as Thompson and Seber suggested. Then, Carfagna and Marzialetti (2009) extended the TSPRN procedure to a sequential setting, introducing an adaptive sequential procedure with permanent random numbers (ASPRN) which allocates one sampling unit at each step. It generates more efficient estimates than the

Thompson and Seber's method and generally than the TSPRN procedure. However, when a cost function with a relevant step cost and budget constraints are considered, the ASPRN may be less efficient than the TSPRN (Carfagna *et al.* (2012)). This result stresses the need of finding the most efficient optimal adaptive sequential procedure, given a cost function. For instance, if the cost per step is high it is more efficient to increase the number of units per step and decrease the number of steps. The optimal sampling procedure depends also on the form of the cost function. Hence, the aim of this paper is to propose an optimal adaptive group sequential procedure with permanent random numbers (optimal AGSPRN) in the presence of a cost function, which is a compromise solution between the ASPRN and the TSPRN procedures. It should preserve the ability of the ASPRN procedure to generate sample allocations very close to Neyman's ones, reducing the impact of the step cost affecting the ASPRN. In Section 2, we describe the AGSPRN procedure and the linear cost function we adopt. In Section 3, we investigate, through a Monte Carlo study, the optimal number of steps K_{opt} and the optimal number of units q_{opt} added at each step characterizing the AGSPRN procedure that generates the estimator with the smallest variance in presence of a linear cost function and budget constraints. The Monte Carlo study requires, as input, information about the distribution function of the analysed population. In some cases, this can be a strong requirement; however it is useful for showing some properties of the optimal AGSPRN procedure and assessing the impact of the components of the cost function. In Section 4, we overcome the limit of the Monte Carlo study proposed in Section 3, setting up a methodology to obtain the optimal AGSPRN procedure when the population is unknown and only a pilot sample is available, that is the usual case. Finally, we discuss the main findings and further developments.

2. The AGSPRN procedure

We propose an adaptive group sequential procedure with permanent random numbers (AGSPRN) that is a group sequential procedure for finite populations which generates a stratified random sample, with adaptive strata allocations at each step. Given a population of size N divided into H strata of size $N_1, \dots, N_H$, for any integer $q \in [1, N - n_0]$, where n_0 is the preliminary or first step sample size, the AGSPRN procedure is developed as following:

(i) assign a random number to each unit in each stratum, then order the units according to the associated number;

(ii) at the first step [$k = 1$] select a first stratified random sample of size n_0 with probability proportional to stratum size, selecting at least two sample units per stratum and estimate the variance inside each stratum;

(iii) compute Neyman's allocation with sample size $n = n_0 + q$ and select q sample units only in the strata with positive difference between Neyman's allocation and the actual one (the allocation is proportional to this difference). Then estimate the parameter of interest and its precision;

(iv) if the stopping rule is satisfied, or the units of the population are all drawn, stop the process; otherwise estimate the strata variances and start again from step (iii) using a sample size equal to $n + q$.

Thanks to the permanent random numbers the selection order of the sample units is assigned at the beginning of the procedure, thus the information gained at step $(K - 1)$ affects only the allocations at step K, not the selection of the units. The allocations are adaptive, not the sample selection.

Let us suppose that we are interested in estimating the population mean $\bar{Y}$ of the some variable Y. A value y_{ih} of Y is associated to each population unit i in stratum h, with $h = 1, \dots, H$. Let K denote the step where the stopping rule is satisfied. The stratified mean estimator in K generated by the AGSPRN procedure that

adds q units at each step is given by:

$$\bar{y}_{stK}(K,q) = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hK} = \sum_{h=1}^{H} \frac{N_h}{N} \sum_{i=1}^{n_{hK}} y_{ihK} ,$$

where h refers to the stratum, $\bar{y}_{hK}$ is the sample mean of stratum h after K steps, n_{hK} is the sample size in stratum h after K steps, y_{ihK} is the value of Y for unit i selected in stratum h after K steps. We are interested in estimating $\bar{y}_{stK}(K,q)$ and its variance:

$$V(\bar{y}_{stK}; K, q) = E\left[\sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_{hK}}{N_h} \frac{\sigma_h^2}{n_{hK}}\right], \quad (1)$$

where σ_h^2 is the variance of Y in stratum h and the expected value is taken with respect to all the possible realizations of the allocations n_{hK} at the Kth step. Because of this complexity, it is prohibitive to compute analytically the value in (1), for each K and each q. Hence, we proceed through a Monte Carlo algorithm that we are going to describe in the next section. Since we don't know the strata variances $\sigma_h^2 s$, for $h = 1, \dots, H$, we are going to estimate them at each step through $S_h^2 = \sum_{i=1}^{n_{hK}} \frac{(y_{ihK} - \bar{y}_{hk})^2}{n_{hK}-1}$, for $h = 1, \dots, H$. Thus, an estimator of $V(\bar{y}_{stK}; K, q)$ is:

$$\hat{V}(\bar{y}_{stK}; K, q) = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \frac{N_h - n_{hK}}{N_h n_{hK}} \sum_{i=1}^{n_{hK}} \frac{(y_{ihK} - \bar{y}_{hk})^2}{n_{hK} - 1}. \quad (2)$$

An important element of our complex framework is the cost. Here we consider a linear cost function:

$$C = C(K,q) = C_0 + c_n[n_0 + q(K-1)] + c_k K, \quad (3)$$

where C_0 is the fixed cost, c_n is the cost per unit, c_k is the cost per step and K is the total number of steps performed by the AGSPRN procedure before stopping.
In this paper we are going to consider the stopping rule regarding the depletion of a budget denoted with C. Our aim is to find the optimal AGSPRN procedure that minimizes the variance of the stratified mean estimator reported in (1), satisfying the cost constraints given by the linear function in (3) for a fixed C.

3. Monte Carlo study

If the pilot sample size n_0, the total budget C, the cost per unit c_n and the cost per step c_k are given, for each $q \in [1, \left\lfloor \frac{(C - C_0 - c_n n_0 - 2c_k)}{c_n} \right\rfloor]$, where $\lfloor \cdot \rfloor$ indicates the integer part rounding to the floor, there is a unique integer K, obtained reversing Equation (3). All these constrained pairs $(K,q)_c = (\left\lfloor \frac{(C - C_0 - c_n n_0 + c_n q)}{c_n q + c_k} \right\rfloor, q)$ belong to the set $\mathcal{H}_{(K,q)_c}$. If two constrained pairs have the same K, we choose the one with the highest q, since the estimator variance is a non increasing function respect to the sample size. As we mentioned, for a fixed q, the distribution of the estimator variance as K increases is analytically intractable. Hence, we proceed through a Monte Carlo study. For each constrained pair $(K,q)_c \in \mathcal{H}_{(K,q)_c}$,

we perform the AGSPRN procedure R times and for each time we estimate the variance of the stratified mean estimator $\hat{V}(\bar{y}_{stK}; K, q)$ as reported in (2). Then we compute the average among the R values of $\hat{V}(\bar{y}_{stK}; K, q)$, obtaining the Monte Carlo estimator variance $\langle \hat{V}^R(\bar{y}_{stK}; K, q) \rangle$ for each constrained pair $(K, q)_c$. The pair that generates the estimator with the lowest Monte Carlo variance characterizes the optimal AGSPRN procedure in presence of budget constraints.

The Monte Carlo algorithm requires some precise knowledge of the entire population in order to perform the AGSPRN procedure for each pair. In the following example we use as input directly the target population, that is chosen to be Normal with different parameters in each stratum. This seems quite useless in a sampling context. However, the aim of this section is to show some properties of the optimal AGSPRN procedure in an ideal situation. In the next section, we are going to estimate the form of the distribution of Y from the pilot sample, updating it at each step in order to provide a precise and useful method for the search of the optimal AGSPRN procedure.

Let us consider the example of a normal distributed population. In each stratum h, for $h = 1, \dots, H$, the variable of interest Y is distributed according to a Normal $\mathcal{N}[h \times 25 + 250, a_h \times 0.9]$, with a_h the h-th element of the vector $a = [600\ 130\ 320\ 250\ 40\ 150\ 100\ 180\ 74\ 400]'$. Let us fix $C = 500, C_0 = 80, c_n = 2, c_k = 4, H = 10, n_0 = 40, W_h = N_h/N \sim \mathcal{U}[450,500]$. The entire population is generated according to this framework, becoming an input of the Monte Carlo algorithm for the search of the optimal AGSPRN procedure. The results are shown in Table 1, where a comparison with TSPRN and ASPRN is also reported, since they are two extreme cases of the AGSPRN procedure with $(K, q) = (2, 166)$ and $(K, q) = (61, 11)$ respectively. As Table 1 shows, the optimal AGSPRN procedure tends not to coincide with TSPRN and ASPRN which generate estimators with higher variance. Hence, our intuition about the existence of a compromise solution between ASPRN and TSPRN that is more efficient in presence of budget constraints and a cost function has been validated. Moreover, a high value of the M. C. estimator variance is also reached using a stratified random sampling (STRS), i.e., drawing all the units obtained with the available budget in just one step with proportional allocation. Using the optimal AGSPRN procedure we gain a variance reduction for the mean estimator of 40% with respect to STRS, with the same budget.

Table 1: Comparison of different adaptive estimators assuming Normal population with C = 500, C_0 = 80, c_n = 2, c_k = 4. The first row presents the optimal solution with the value of $\bar{y}_{stK}$, its variance, the MCE, the sample size n and the pilot size n_0. The consecutive rows show the comparisons with other sampling procedures: TSPRN, ASPRN and STRS. Here, $\bar{Y}$ = 391.35.

	K	q	$\bar{y}_{stK}$	$\langle \hat{V}^R(\bar{y}_{stK}; K, q) \rangle$	MCE	n	n_0
Optimal AGSPRN	4	54	391.53	175.80	0.221	202	40
TSPRN	2	166	391.71	195.43	0.233	206	40
ASPRN	57	1	391.84	365.76	0.500	96	40
STRS	1	0	391.72	297.04	0.300	207	207

Table 2 shows the impact of different values of the cost components on the optimal AGSPRN procedure. For instance, it is possible to notice that an increase of c_n under a fixed budget causes an increase of K_{opt} and a decrease of q_{opt}, with a consequent decrement of the total sample size and a negative impact on the estimator variance that becomes higher. The number of steps K_{opt} tends to be maintained high, since a decrease of K_{opt} inflates the estimator variance relatively more than a decrease of q_{opt}. On the other hand, an increase of c_k generates a decrease of K_{opt}, but not of the total sample size, with a lower effect on the variance of the estimator.

Table 2: Effect of the cost components for normally distributed data in presence of budget constraints.

c_n	c_k	K_{opt}	q_{opt}	$\bar{y}_{stK}$	$\langle \hat{V}^R(\bar{y}_{stK}; K, q) \rangle$	MCE	n
2	4	4	54	391.53	175.80	0.221	202
2.5	4	5	30	391.42	221.07	0.277	160
3	4	5	23	391.54	268.12	0.336	132
4	4	5	15	391.87	357.63	0.473	100
c_n	c_k	K_{opt}	q_{opt}	$\bar{y}_{stK}$	$\langle \hat{V}^R(\bar{y}_{stK}; K, q) \rangle$	MCE	n
2	2	5	41	391.59	173.05	0.216	204
2	4	4	54	391.53	175.80	0.221	202
2	6	3	80	391.52	180.23	0.219	200
2	8	3	79	391.57	181.87	0.222	198

4. The search of the optimal AGSPRN procedure when the target population is unknown

In this section we propose a method for the search of the optimal AGSPRN procedure when the distribution of Y is unknown and it is estimated at each step k through kernel techniques or model assumptions with estimated parameters in each stratum. This is an extension of the bootstrap method of Rosenberger and Hu (1999) who applied it to infinite populations with Bernoulli distribution in the clinical trials context. Our proposal for the search of the optimal AGSPRN procedure is developed in the following steps:

(i) at the first step $[k = 1]$ select a first stratified random sample of size n_0 with proportional allocation, selecting at least two sample units per stratum and estimate the variance inside each stratum;

(ii) make some assumptions on the distribution form of Y inside each stratum using the selected units, estimate the parameters and generate from that distribution $N_h - \lfloor n_0^h \rfloor$ values, $h = 1, \dots, H$, such that all the finite population of size N is obtained; here $\lfloor n_0^h \rfloor$ is the integer part rounding to the floor of the n_0 units allocated to stratum h;

(iii) using the estimated population, simulate R times the AGSPRN procedure with different values of q and choose the optimal pair (K_{opt}, q_{opt}) that minimizes the estimator variance given budget constraints;

(iv) compute Neyman's allocation with sample size $n = n_0 + q_{opt}$ and select q_{opt} sample units only in the strata with positive difference between Neyman's allocation and the actual one (the allocation is proportional to this difference). Then estimate the parameter of interest and its precision;

(v) if the stopping rule is satisfied and $K_{opt} = 2$ stop the process, otherwise start again from step (ii) fixing $n_0 = n_0 + q_{opt}$ and $C_0 = C_0 + c_k$.

The optimal AGSPRN procedure obtained through this method is characterized by a number of units added at each step that can vary from step to step, depending on the updated population. This procedure allows to obtain sample allocations as close as possible to Neyman's ones, by computing, at each step, the allocations and the combination of number of steps and of units per step, given the cost function in (3).

We apply the described method using the same framework of the previous section. We select a pilot sample from the population generated by the Normal distributions and we get the same result of Table 1: the optimal AGSPRN consists of 4 steps and 54 units per step. The results obtained by estimating the population step

by step and those generated in the ideal situation of a known population coincide: this confirms the validity of the method proposed in this section.

5. Discussion

The need of filling the gap regarding an adaptive sequential sampling procedure for finite populations that is optimal in terms of minimum variance of the estimator given budget constraints and a cost function has lead us to propose an adaptive group sequential procedure with permanent random number (AGSPRN) for stratified finite populations. It has been shown to be more efficient than other existing sampling procedures when the step cost is relevant. Moreover, it tends not to coincide with the two steps adaptive procedure with permanent random numbers (TSPRN) and the adaptive sequential procedure with permanent random numbers (ASPRN) that can be derived as two particular cases.
In Section 3 we have assessed the impact of the cost components on the optimal AGSPRN procedure, finding out that an increase of the step cost and consequently a decrease of the optimal number of steps inflate the estimator variance relatively more than an increase of the cost per unit. Only a linear cost function has been considered, but the impact of other kinds of cost function on the optimal procedure could be assessed in a future research. In Section 4 we have proposed a method to obtain the optimal AGSPRN procedure when the distribution of the variable of interest is unknown. It performs quite well, leading to the same results obtained in the ideal situation of a known population. Moreover, it adds flexibility to the optimal procedure since it allows the number of units added at each step to vary from step to step.
Other criteria of stopping should be investigated and the study can be conducted for more than one variable of interest, considering multiple adaptive allocation methods.

References

Carfagna E. (2007). Crop area estimates with area frames in the presence of measurement errors. In Proceedings of the Fourth International Conference on Agricultural Statistics. Advancing Statistical Integration and Analysis, 22- 24 October 2007. 1–10. Beijing.

Carfagna E., Marzialetti J. (2009). Sequential design in quality control and validation of land cover data bases. Journal of Applied Stochastic Models in Business and Industry (ASMBI) 25-2: 195–205.

Carfagna, E., Tassinari, P., Zagoraiou, M., Benni, S., Torreggiani, D. (2012). Efficient statistical sample designs in a GIS for monitoring the landscape changes. In Advanced Statistical Methods for the Analysis of Large Data-Sets edited by Di Ciaccio et al., 399–407. Springer, Berlin Heidelberg.

Chow, Y.S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. The Annals of Mathematical Statistics, 36, pp. 457–462.

Ray, W.D (1957). Sequential confidence intervals for the mean of a normal distribution with unknown variance. Journal of the Royal Statistical Society, Ser. B. 19, 133–143.

Stein, C. (1949). Some problems in sequential estimation. Econometrica. 17, 77–78.

Ohlsson, E. 1995. Coordination of samples using permanent random numbers. In Business Survey Methods edited by Cox, B. et al., 153-169. New York: Wiley.

W. F. Rosenberger and F. Hu (1999). Bootstrap methods for adaptive designs. Statistics in Medicine, 18(14):1757–1767.

Thompson, S.K., Seber, G.A.F. (1996). Adaptive Sampling. New York: Wiley.