

Two-Stage Adaptive Sampling

Campionamento Adattivo a due stadi

Stefania Naddeo

Università di Siena

Dipartimento di Metodi Quantitativi

naddeo@unisi.it

Caterina Pisani

Università di Firenze

Dipartimento di Statistica

pisani@ds.unifi.it

Riassunto: Sebbene il campionamento adattivo venga frequentemente applicato per la stima delle abbondanze di popolazioni biologiche con comportamento spaziale di tipo contagioso, esso si basa sull'assunzione che tutti gli individui presenti all'interno delle unità selezionate siano osservati. Dal momento che tale assunzione risulta irrealistica nel caso di popolazioni elusive, il lavoro analizza le proprietà di una procedura adattiva a due stadi in cui le abbondanze all'interno di ogni unità selezionata sono stimate invece di essere rilevate senza errore.

Keywords: adaptive cluster sampling, two-stage estimation, replicated counts.

1. Introduction

Adaptive sampling offers a suitable solution to the problem of estimating the abundance of rare, clustered populations. The design involves the selection of an initial sample of area units and neighbouring units are added whenever a given number of individuals is recorded within. It is worth noting that adaptive sampling is based on the assumption that every member of the population in the selected units is observed. In many real situations this assumption may be unrealistic, such as when dealing with fishes or hardly detectable animals, even if inference is usually made as detectability was perfect. The aim of this paper is to check the performance of adaptive sampling in the more realistic situation in which abundance in the selected units is estimated instead of being recorded without errors. To this purpose, a two-stage adaptive sampling design is considered. In the first stage an initial sample of units is selected by simple random sampling without replacement, while the second stage involves the estimation of abundance within the sampled units by using an encounter strategy. Accordingly, the total number of units included in the final sample depends on the values of the resulting estimates. The statistical properties of the two-stage estimator are derived and a simulation study is carried out to check the performance of this strategy.

2. Statistical properties

Consider a study region partitioned in N spatial units and denote by $T_1, T_2, \dots, T_N$ the unit abundance. Let T be the whole abundance over the study area. An initial sample of n units is selected by simple random sample without replacement. If $\{1, 2, \dots, N\}$ denotes the set of indexes labelling the population units then the initial sample may be viewed as

a set of indexes $S_0 \subset \{1, 2, \dots, N\}$. Now the abundance in each selected unit $j \in S_0$ is not observed but it is estimated by means of an encounter sampling strategy. If the encounter procedure is independently replicated m_j times and the corresponding Horvitz-Thompson abundance estimate is computed, then the m_j replications give rise to m_j iid random variables with expectation T_j and variance s_j^2 . Then the average of the m_j estimates, say $\hat{T}_j$, constitutes the realization of a random variable with expectation T_j and variance s_j^2 / m_j . Moreover $\hat{T}_j$ is asymptotically normal ($m_j \rightarrow \infty$) and an unbiased estimator of its variance is trivially obtained by means of s_j^2 / m_j , where s_j^2 is the sample variance of the m_j estimates. Whenever $\hat{T}_j$ satisfies a given condition $\hat{T}_j \in C$ (e.g. $\hat{T}_j > 0$), additional units in the neighbourhood of the j -th unit are added to the sample. For each additional unit h , if $\hat{T}_h \in C$ the neighbouring units are also observed, and so on until a final sample S is obtained. Note that S is composed by clusters of units, each one formed by a boundary of units in which the estimate does not satisfy the condition (the so-called edge units) and by a network of units whose estimates belong to C . As suggested by Thompson and Seber (1996), it is convenient to consider any unit not satisfying C a network of size one, so that the population may be uniquely partitioned into networks. It is worth noting that the partition is random depending on the random vector $\hat{\mathbf{T}} = [\hat{T}_1, \hat{T}_2, \dots, \hat{T}_N]^T$. If K is the number of networks in the population, R is the set of indexes labelling the networks intersected by the initial sample and S_k denotes the set of the n_k indexes labelling the units in the k -th network, all conditioned to $\hat{\mathbf{T}}$, then the probability that the initial sample intersects the k -th network ($k \in R$) turns out to be

$$a_k = 1 - \frac{\binom{N - n_k}{n}}{\binom{N}{n}}.$$

Accordingly, the two-stage estimator of total is

$$\hat{\hat{T}} = \sum_{k \in R} \frac{\hat{T}_k^*}{a_k}, \quad (1)$$

where

$$\hat{T}_k^* = \sum_{j \in S_k} \hat{T}_j$$

constitutes an unbiased estimator of the k -th network abundance, say T_k^* , with variance

$$s_k^{*2} = \sum_{j \in S_k} \frac{s_j^2}{m_j}$$

which can be unbiasedly estimate by

$$s_k^{*2} = \sum_{j \in S_k} \frac{s_j^2}{m_j}$$

As to the properties of the two-stage estimator (1), it is at once apparent that

$$\begin{aligned} E(\hat{T}) &= E_{\hat{\mathbf{T}}} \left[E_R(\hat{T} | \hat{\mathbf{T}}) \right] = E_{\hat{\mathbf{T}}} \left[\sum_{k=1}^K \hat{T}_k^* \right] = T, \\ V(\hat{T}) &= V_{\hat{\mathbf{T}}} \left[E_R(\hat{T} | \hat{\mathbf{T}}) \right] + E_{\hat{\mathbf{T}}} \left[V_R(\hat{T} | \hat{\mathbf{T}}) \right] = \\ &= \sum_{k=1}^K s_k^{*2} + E_{\hat{\mathbf{T}}} \left[\sum_{k=1}^K \frac{(\hat{T}_k^*)^2 (1 - a_k)}{a_k} + 2 \sum_{k=1}^{K-1} \sum_{h=j+1}^K \hat{T}_k^* \hat{T}_h^* \frac{(a_{kh} - a_k a_h)}{a_k a_h} \right], \end{aligned} \quad (2)$$

where

$$a_{kh} = 1 - \frac{\binom{N-n_k}{n} + \binom{N-n_h}{n} - \binom{N-n_k-n_h}{n}}{\binom{N}{n}}$$

is the probability that the initial sample intersects both the k -th and the h -th network conditioned on $\hat{\mathbf{T}}$. Note that (2) differs from the variance of the classical adaptive estimator (Thompson and Seber, 1996). The first term depends on the estimation within the selected units while the second term depends on both the selection and the estimation procedure and cannot be developed further. However, an unbiased estimator of (2) may be straightforwardly obtained by

$$\hat{V} = \sum_{k \in R} \frac{s_k^{*2}}{a_k} + \sum_{k \in R} \frac{(\hat{T}_k^*)^2 (1 - a_k)}{a_k^2} + 2 \sum_{k \in R} \sum_{h > k} \hat{T}_k^* \hat{T}_h^* \frac{(a_{kh} - a_k a_h)}{a_k a_h a_{kh}}$$

3. Some Monte Carlo simulations

The artificial population of $N=400$ squared units of size one described by Thompson (1992, p.285) is considered and the individuals in each unit have been allocated according to a regular process. In order to check the performance of the two-stage

procedure, 10,000 initial samples of size $n=5(5)20$ units were selected by simple random sampling without replacement. For each selected sample both the classical adaptive cluster strategy as well as the two-stage strategy are performed. As to the two-stage strategy, the abundance within each selected unit is estimated by a plot sampling procedure performed by using $m_j=10(10)30$ circular plots with radius $r=0.06$.

On the basis of the 10,000 samples, the empirical variances (EV) of the two-stage abundance estimator is computed together with the empirical expectations of the sample sizes (ESS) arising from the adaptive procedure. Moreover, in order to compare the performance of the two-stage procedure with respect to the classical one, the relative increase in variability (RIV) due to estimation within the sampled units is computed together with the average relative decrease of the effective surveyed surface (RDS). Results of the simulation, together with the exact values of the variance (V) of the classical estimator are reported in Table 1.

Table 1: Empirical comparison of two-stage adaptive procedure with respect to the classical adaptive procedure.

n	number of plots	Two-stage Adaptive Procedure		Classical Adaptive Procedure		RIV	RDS
		EV	ESS	V	ESS		
5	10	2,644,895.33	0.58	131,257.92	9.40	19.15	0.94
	20	1,408,257.42	1.18	131,257.92	9.33	9.73	0.87
	30	1,090,103.64	1.83	131,257.92	9.56	7.30	0.81
10	10	1,799,755.19	1.15	61,848.27	18.41	28.10	0.94
	20	817,594.05	2.36	61,848.27	18.36	12.22	0.87
	30	463,358.77	3.63	61,848.27	18.13	6.49	0.80
15	10	1,056,190.22	1.73	38,805.13	26.90	26.22	0.94
	20	530,972.40	3.55	38,805.13	26.65	12.68	0.87
	30	308,223.99	5.44	38,805.13	26.64	6.94	0.80
20	10	756,059.64	2.31	27,355.20	34.73	26.64	0.93
	20	405,427.28	4.72	27,355.20	34.63	13.82	0.86
	30	248,325.45	7.24	27,355.20	34.62	8.08	0.79

4. Concluding remarks

From the previous results it is at once apparent that the variance of the two-stage estimator is dramatically higher than the variance of the classical one and the increase in variability can not be explained even taking into account the decrease in the effective surveyed surface. Accordingly, in absence of a perfect detectability, the increase of the variance due to the estimation within the selected units cannot be neglected in order to avoid unreliable evaluations of the precision of the resulting estimates.

References

- Thompson S. K. (1992) *Sampling*, John Wiley and Sons..
Thompson S. K., Seber, G. A. F. (1996) *Adaptive Sampling*, John Wiley and Sons.