Bayesian Population Size Estimation with A Single Sample

Stima della numerosità di una popolazione utilizzando un solo campione

Pierfrancesco Alaimo Di Loro and Luca Tardella

Abstract The estimation of the size of a finite population is a problem encountered in a variety of applications. One standard statistical approach relies on *mark-recapture* sampling, which may require high costs and annoyance to the population of interest. These considerations have motivated the search for alternative sampling strategies that allow to estimate the size of a population from a single capture. Hettiarachchige [4] proposes a method that is viable when the population is made of only two generations: a group of generators and one of generated units. We investigate Bayesian methods alternative to the frequentist estimators used in [4]. Preliminary results give evidence of competing performance of the Bayesian approach, which in some cases sensibly outperforms the frequentist alternatives.

Abstract La stima della numerosità di una popolazione è un problema comune a vari ambiti di applicazione. Le procedure di stima sono solitamente basate sul noto metodo *cattura-ricattura*, il quale comporta elevati costi e disturbo della popolazione. Tali considerazioni hanno stimolato la ricerca di tecniche che permettano di ottenere un stima utilizzando un unico campione. Hettiarachchige [4] propone un metodo applicabile nel caso in cui la popolazione sia composta di due sole generazioni: un gruppo di unità generatrici ed uno di unità generate. L'obiettivo del nostro lavoro è quello di ottenere un'estensione Bayesiana dell'originale modello frequentista. Risultati preliminari evidenziano accuratezza degli stimatori Bayesiani sensibilmente migliore rispetto alle alternative frequentiste.

Key words: population size, capture-recapture, mark-recapture, single sample, Bayesian inference, moment estimator, MLE.

Pierfrancesco Alaimo Di Loro
University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185 Roma RM, e-mail: pierfrancesco.alaimodiloro@uniroma1.it

Luca Tardella
University of Rome "La Sapienza", Statistical Science Department, Piazzale Aldo Moro 5, 00185 Roma RM, e-mail: luca.tardella@uniroma1.it

1 Introduction

The problem of estimating the size or any other demographic parameter of a population of interest for which there is no complete enumeration or reference list is common to a variety of applications: ecology (e.g. natural and wildlife populations), reliability, epidemiology, social sciences. However, most of the literature regarding this matter has been developed in the statistical ecology field, where *capture-recapture* methods have been the ruling paradigm for the whole second half of the 20th century.

The modern foundation of these methods was laid in [2] and [6] and they are all based on the pioneering *mark-recapture* sampling technique which originated the well-known *Lincoln-Petersen* estimator. The most basic version consists of taking a random sample of size n_1 from the population and mark the captured individuals. They are then returned to the population and, at a later occasion, a second sample of size n_2 is taken. The previously applied tags allow to recognize if and how many of the captured individuals were already been sampled at the previous occasion. If m of them already have a tag, then the *Petersen* estimator is: $\hat{N}_P = (n_1 n_2)/m$.

The biggest issue with the application of such methods is that they require the population to be sampled at least twice. The necessity of at least one further capture occasion leads to increasing costs and, furthermore, can cause an ever-increasing annoyance to the population of interest. The latter can alter its natural equilibrium, leading to a change of conditions from one capture to the other. This may introduce bias in the estimates when those changes are not taken into proper account and requires behavioural adjustments on the basic model [3] . Moreover, there are a lot of situations in which the captured individuals cannot be returned to the population, making the procedure impractical. These considerations have motivated the search for procedures that allow to estimate the size of a population in alternative ways.

In the last decades genetic data have become increasingly important in ecology and conservation biology and their use in estimating the population size have been considered [5]. The underlying idea is that the degree of biological relationship between a sample of individuals from the population provides information about the population itself and DNA profiles can be used to detect the degree of relatedness between individuals. [7] exploit this idea in analogy to the traditional capture-recapture method and argued in favor of a *single sample* version of the Petersen estimator. An individual is marked by its presence in the sample, and "recaptured" if the sample contains one or more close relatives. In practice, it allows to generalize from "recapture of self" to "recapture of closely-related kin" ([1]), where the detection of kinship is based on the idea that an individual share more alleles with a parent than with a biologically unrelated individual.

However, as it has already been underlined in [7], this approach is sensibly more complex than the ordinary *capture-recapture* method. The effectiveness of this method depends on hypotheses on the population that are hardly matched in real life and relies on accurate estimates of the kinship coefficients between individuals. The complexity of the method has been reduced by [4], who put himself in a

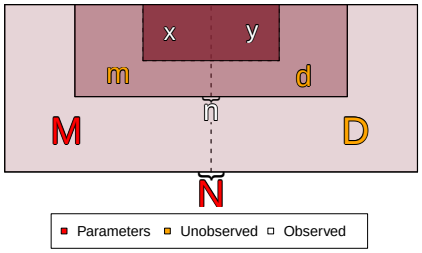

Fig. 1 Schematic visualization of the structure of population and of the quantities involved in the sample. Different colors are used to distinguish between parameters, latent variables and observed quantities.

simpler framework where kinship misclassification plays a slightly less important role. The framework of study will be more extensively investigated in Section 2.

Our contribution to [4]'s work is to introduce a suitable way of implementing Bayesian methods (Section 3) alternative to the frequentist estimators used by the original author. A comparative analysis of these estimators is provided at the end of the section. We finally provide an outline of some promising developments and extensions that may improve on the precision of the proposed estimates in Section 4.

2 General Framework

Let us introduce formally the assumptions of the model. The population structure considered in [4] is composed of only two generations: the individuals from the first generation are denoted as *mothers* and the individual from the second as *daughters*. For the purpose of this paper our main interest will be in estimating the size of the population of the *mothers*.

The two generations are assumed mutually exclusive and collectively exhaustive, and the population to be closed. A sample is taken and perfect identifiability of mother-daughter couples is assumed. The appropriateness of this assumption is discussed in [4].

In practice, we are dealing with a random sample from a population of N individuals, where M are "mothers" and D are "daughters" ($N = M + D$). The captured individuals will be in part mothers and in part daughters. We are able to recognize

Latent		**Observables**	
		$n = m + d$	all units in the sample
m	all mothers in the sample	x	all *identified* mothers in the sample
$(d_i)_{i=1}^{m}$	daughters for each mother in the sample	$(y_i)_{i=1}^{x}$	*id.* daughters for each *id.* mother in the sample
$d = \sum_{i=1}^{m} d_i$	all daughters in the sample	$y = \sum_{i=1}^{x} y_i$	all *identified* daughters in the sample

Table 1 Quantities of interests involved in the single sample

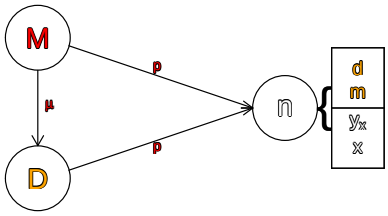

Fig. 2 Graphical visualization of the generating process of the data. Parameters in *red*, latent variables in *orange* and observables in *white*.

as daughters only those daughters whose mothers have been captured and viceversa. The relevant quantities involved in the single sample are visualized in Figure 1 and listed in Table 1 with appropriate notation:

Probabilistic Model. It is assumed that each mother has generated, at a previous time, a certain number of daughters D_i according to a $Pois(\mu)$. The total number of daughters in the population is then $D = \sum_{i=1}^{M} D_i \sim Pois(M\mu)$. The parameter μ is constrained to the set $[1, +\infty)$ by [4] for reasons related to the existence of the moment estimator. Our Bayesian approach, theoretically, does not require such an assumption but we will stick to this constraint to ensure a fair comparison. Furthermore, each individual is supposed to be captured independently with equal probability p.

The total number of mothers in the population M is the parameter of interest, while μ and p are just nuisance parameters. The situation is graphically reported in Figure 2.

An explicit form of the *marginal* likelihood can be obtained using a conditioning argument:

$$P(n, x, (y_i)_{i=1}^{x} | M, \mu, p) = \sum_{m=x}^{M \wedge (n-y)} P(n, x, (y_i)_{i=1}^{x} | m, M, \mu, p) P(m | M, p). \tag{1}$$

$P(m|M,p)$ is the probability to capture independently m mothers given that there are M mothers in the population and they are captured with probability p[1], which is:

$$P(m|M,p) = Bin(m|M,p) = \binom{M}{m} p^m (1-p)^{M-m}.$$

The joint density of $(n, x, (y_i)_{i=1}^{x})$ conditioned on (m, M, μ, p), and hence the likelihood of the model, can be shown to be equal to:

$$P\left(n, x, (y_i)_{i=1}^{x} | \, m, M, \mu, p\right) = \binom{m}{x} \frac{e^{-M\mu p} (\mu p)^y}{\prod_{i=1}^{x} y_i!} \frac{((M-m)\mu p)^{(n-m-y)}}{(n-m-d)!}.$$

[1] This probability obviously does not depend on the mean number of daughters μ per mother.

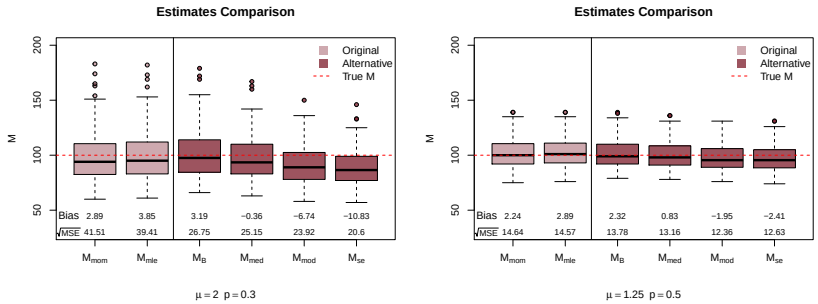

Fig. 3 Box-plots of the resulting estimates obtained for the 100 different simulation set for each estimator. From left to right: *Moment* and *Maximum Likelihood* estimators, posterior *Mean*, *Median*, *Mode* and *PMSE minimizer*

3 The Bayesian Extension

Some undesirable peculiarity of the likelihood such as presence of non-unique solutions and computational problems led [4] to discard the maximum likelihood estimator. The author decided to resort to a moment-based estimator in order to overcome these problems. However, the method of moments is still affected by well-known boundary solution instability.

We propose an estimation procedure based on a Bayesian approach, which should be able to regularize the model likelihood without incurring in the moment estimator deficiencies. Furthermore, Bayesian methods allow to include prior information on the demographic parameters of the population whenever such information is available and can eventually replace in a sensible way the *ad-hoc* constraint on μ.

Independent priors have been assigned to the parameters M, μ and p, so that the posterior distribution can be written as:

$$\pi(M,\mu,p|n,x,(y_i)_{i=1}^x) \propto \mathscr{L}(M,\mu,p;n,x,(y_i)_{i=1}^x)\pi(M)\pi(\mu)\pi(p)$$

The forms chosen for the priors of each parameter are:

- low-informative $TruncatedGamma(0.05,0.025)$ and $Beta(0.001,0.001)$ priors for μ and p;
- inverse prior $\pi(M) \propto \frac{1}{M^a}$ for M, with $a = 0,1,2$.

Posterior samples from the joint posterior of the parameters have been obtained via *Metropolis-Within-Gibbs* algorithm, where a Gibbs-style update is performed for M and a bi-variate Normal random walk M-H is performed for (μ,p). The considered Bayesian estimators are: the posterior mean $\hat{M}_B$, the posterior median $\hat{M}_m$ and the approximated PMSE minimizer[2] $\hat{M}_{se}$.

[2] The *Posterior Mean Squared Error* minimizer is the value that minimizes the MSE with respect to the posterior distribution of the quantity of interest: $\text{argmin}_{a\in\mathscr{M}} \sum_{M\in\mathscr{M}}(M-a)^2\pi(M|\cdot)$, where M is the quantity of interest and $\mathscr{M}$ its domain.

Preliminary Results. A simulation study has been carried out in order to verify the effectiveness of the proposed Bayesian estimators. For different configurations of μ and p, with fixed $M = 100$ mothers, the simulation of mother-daughter sampling have been replicated for $S = 100$ times, producing $S = 100$ realizations of all the alternative estimators. The comparative performance is assessed in terms of *Mean Squared Error*. Our preliminary results are exposed in Figure 3 and give evidence of competing performance of the Bayesian approach which, especially in some configurations, sensibly outperforms the frequentist alternatives.

4 Concluding remarks and further Developments

We have proposed a Bayesian framework for the estimation of the population size in presence of a single sample. This technique relies on the pairing of mothers and daughters in the sample through the use of genetic markers on the line of [4]. We have shown that the Bayesian framework allows to reduce the error of the classical estimators up to 50% in specific parameter settings. Indeed, we are currently workin on many other improvements and extensions to the proposed Bayesian methodology:

1. formal derivation of non-informative priors and principled informative priors possibly removing the unnatural *ad-hoc* constraint $\mu > 1$;
2. inclusion of other kind of kinships and/or other covariates in order to reduce the variability of the unidentified part of the sample;
3. relaxation of restrictive model assumptions like the identical capture probability and the perfect identification.

References

1. Bravington, M.V. and Skaug, H.J. and Eric, C.: Close-kin mark-recapture. Statistical Science 31 (2), 259 – 274 (2016), Institute of Mathematical Statistics
2. Cormack, R.M.: Estimates of survival from the sighting of marked animals. Biometrika 51 (3/4), 429 - 438 (1964), JSTOR
3. Fegatelli, D.A. and Tardella, L.: Improved inference on capture recapture models with behavioural effects. Statistical Methods & Applications 22 (1), 45 - 66 (2013), Springer
4. Hettiarachchige, C.K.H.: Inference from single occasion capture experiments using genetic markers. PhD Thesis (2016)
5. Schwartz, M.K. and Tallmon, D.A. and Luikart, G.: Review of DNA-based census and effective population size estimators. Animal Conservation forum 1 (4), 293 - 299 (1998), Cambridge University Press
6. Seber, G.A.F.: A note on the multiple-recapture census. Biometrika 52 (1/2), 249 - 259, (1965), JSTOR
7. Skaug, H.J.: Allele-sharing methods for estimation of population size. Biometrics 57 (3), 750 – 756 (2001), JSTOR