

Optimal predictive sample size determination for interval estimation

Numerosità campionaria ottima in problemi di stima per intervallo

Fulvio De Santis

Dipartimento di Statistica, Probabilità
e Statistiche Applicate
Università di Roma “La Sapienza”
fulvio.desantis@uniroma1.it

Marco Perone Pacifico

Dipartimento di Statistica, Probabilità
e Statistiche Applicate
Università di Roma “La Sapienza”
marco.peronepacifico@uniroma1.it

Riassunto: Si affronta il problema della scelta della numerosità campionaria in problemi di stima per intervalli. Il criterio proposto, sviluppato in ambito bayesiano, consiste nel limitare la probabilità di ottenere intervalli di credibilità di ampiezza superiore ad una soglia prefissata. Rispetto alle procedure disponibili, basate sulla lunghezza media degli intervalli, il criterio considerato permette un controllo anche della variabilità di tale lunghezza. Vengono determinati risultati analitici per problemi di stima della media (riportati nel lavoro) e della differenza tra medie di popolazioni normali. Si propone inoltre una procedura numerica che permette di trattare situazioni in cui non è facile ottenere risultati in forma chiusa. Il criterio e la procedura numerica vengono illustrati attraverso la loro applicazione a problemi biomedici, in particolare ad analisi di sopravvivenza e studi caso-controllo.

Keywords: Bayesian design; case-control studies; censoring; credible interval; interval estimation; odds-ratio; sample size; survival.

1. Introduction and motivation

Determination of the optimal sample size is a fundamental step in many statistical designs, especially relevant in biomedical studies and in all situations where data are difficult to collect.

The literature regarding sample size choice and, more generally, optimal design is quite large, especially within the frequentist inferential framework. The most common frequentist criteria for sample size determination are related to power considerations when the statistical problem is testing and, in estimation problems, to either size of intervals or variance of estimators. Typically, standard frequentist criteria for sample size choice depend on one or more unknown nuisance parameters, so that for using these procedures initial guesses on the true values of these parameters are needed. For instance, optimal sample size formulas for testing a normal mean depend on the variance that must be replaced by arbitrary guesses. Hence the resulting sample sizes are only locally optimal and can be quite sensitive to the values chosen.

In the Bayesian approach this problem is automatically overcome by using a prior distribution for all the unknown parameters in the model. The prior distribution allows the statistician to use all the information on the parameters of the model, to take into account their uncertainty and to integrate out all the nuisance parameters. This results in greater flexibility than the classical approach.

Although the literature on Bayesian sample size determination is not as large as the

frequentist one, the problem has been considered by several authors (see, among the others, Raiffa and Schlaifer (1961), Piccinato (1996), Lindley (1997)). For an exhaustive overview on Bayesian sample size choice see Adcock (1997) and on Bayesian experimental design see Chaloner and Verdinelli (1995).

Joseph and Belisle (1997) and Joseph *et al.* (1997) propose a criterion, that we review in Section 2, based on the expected length of credible sets. In order to take into account the variability of the random interval length, we propose here a method based on the probability of having credible sets with length not greater than a fixed threshold ℓ_0 . Analytic results are provided for the normal mean problem. Unlike the normal case, when the posterior density is not symmetric, finding analytic results on credible sets might be difficult, so we provide a numerical method to determine the sample size according to the criteria considered.

2. Problem setting

We are interested in interval estimation of a one-dimensional parameter of interest θ . We want to find the optimal size n of a random sample $\mathbf{X}_n = (X_1, \dots, X_n)$, whose joint density depends on some unknown parameters.

The most common Bayesian interval estimates for θ are equal-tails and highest posterior density (HPD) intervals. From the computational point of view, equal-tails are much easier to work with than HPD's. This reason and the fact that HPD sets are not invariant under parameter transformations make the use of equal-tails intervals very common. Conversely, HPD intervals are of minimal size among all the sets with posterior probability equal to $1 - \alpha$ and this renders HPD sets recommendable whenever a "preferred" parametrization can be identified. Note that HPD and equal-tails coincide when the posterior density is symmetric. We will focus our presentation on HPD sets, but it can as well be rephrased in terms of equal-tails intervals.

In sample size determination problems the idea is that, before performing the experiment yielding the data, we want to choose the minimal sample size that satisfies a selected criterion. We are then in a pre-experimental setting and both the interval estimator and its length are functions of the random sample $\mathbf{X}_n$. Thus all pre-experimental computations should be done keeping the randomness of $\mathbf{X}_n$ into consideration. In the following we will denote with $\mathbb{P}$ the predictive distribution of the random sample $\mathbf{X}_n$ and with $\mathbb{E}$ the expectations computed through it.

Among the several criteria proposed in the Bayesian literature we will focus on the so-called *average length criterion* (ALC), also examined, for instance, by Joseph and Belisle (1997) and Joseph *et al.* (1997). The idea is to find, for a fixed α , the smallest sample size such that the expected length of the $(1 - \alpha)$ -level HPD is less than a chosen value, ℓ . Then, denoting by $L_\alpha(\mathbf{X}_n)$ the random variable corresponding to the length of the $(1 - \alpha)$ -level HPD, we look for the smallest n such that

$$\mathbb{E}[L_\alpha(\mathbf{X}_n)] \leq \ell, \quad (1)$$

where the expectation is with respect to the marginal distribution of the sample $\mathbf{X}_n$. Therefore, this criterion only requires an average control on the length of the HPD set, with no control on the variability of its size.

The basic idea leading one to choose ALC is to reduce, before the experiment is performed, the probability of having large values for $L_\alpha(\mathbf{X}_n)$. Actually, when (1) is

satisfied, using Markov inequality we obtain

$$\mathbb{P}[L_\alpha(\mathbf{X}_n) \geq \ell_0] \leq \frac{\ell}{\ell_0}$$

which is a coarse bound that would lead one to choose a sample size larger than needed.

In view of the above considerations, we propose the *length probability criterion* (LPC): we select n in order that the probability of having an HPD for θ with length greater than ℓ_0 is less than a chosen level γ . Then, we choose the smallest n such that

$$\mathbb{P}[L_\alpha(\mathbf{X}_n) \geq \ell_0] \leq \gamma. \quad (2)$$

Unlike ALC, criterion (2) allows one to control also the variability of the random length of credible sets.

Note that, under rather mild conditions, as $n \rightarrow \infty$ the posterior distribution of θ tends almost surely to a point mass, so that $L_\alpha(\mathbf{X}_n) \rightarrow 0$. This ensures the existence of some n satisfying the LPC; moreover, asymptotic normality makes the use of HPD and equal-tails equivalent for large samples.

Besides sample size choice, plotting $\mathbb{P}[L_\alpha(\mathbf{X}_n) \geq \ell_0]$ as a function of n can be extremely useful in order to visualize whether an increment in the sample size reduces significantly the probability of having large set estimates.

3. Implementation of the criteria

For dealing with the criteria previously introduced, one needs to study the sampling distribution of the length of the HPD set. When the posterior density is symmetric, HPD and equal-tails sets coincide and determining the sampling distribution of $L_\alpha(\mathbf{X}_n)$ might be feasible. For instance it can be shown that, when the parameter of interest is the mean of a normal distribution with unknown variance, the probability in (2) is

$$\mathbb{P}[L_\alpha(\mathbf{X}_n) \geq \ell_0] = \int_0^u \text{Be}(t|\nu, \frac{n}{2})dt$$

where $\text{Be}(\cdot|\alpha, \beta)$ denotes the beta density and u is a function of n, ℓ_0 and of the conjugate prior parameters. A similar closed-form expression is obtained when the parameter of interest is the difference between the means of two normal distributions. This problem is especially relevant as it covers many practical situations, like comparing the effects of two treatments in clinical trials.

When the posterior density is not symmetric, finding an analytic expression for the HPD set is not easy and studying the sampling distribution of $L_\alpha(\mathbf{X}_n)$ turns out to be awkward. In order to use our criterion for determining the sample size, we develop a simple numerical procedure to approximate HPD sets and to simulate the distribution of the length $L_\alpha(\mathbf{X}_n)$. The procedure makes it possible to use LPC in very general contexts since it only requires sampling from the marginal (which is always straightforward) and from the posterior (which is affordable, even in the non-conjugate case, using MCMC techniques). Accuracy of the numerical procedure has been tested comparing numerical and analytical results in the normal case.

The criterion introduced above has been used in relevant applied biostatistical contexts as survival analysis and case-control studies. In these problems conjugate prior

distributions have been used and prior parameters have been elicited using information from previous studies.

In a survival analysis study, aimed to estimate the mean survival time for acute leukemia patients submitted to 6-Mercaptopurine therapy, LPC provides a tool useful both for determining the optimal sample size and for evaluating, before the experiment is performed, the impact of different censoring times (and hence of censoring fractions) on the analysis. In this study prior elicitation was based on results from Freireich (1963).

In the case-control setting the criterion is used to choose the sample size and to determine the optimal allocation of units in order to plan a study on the association between non-Hodgkin Lymphoma and exposure to pesticides. In this case the source of information for prior elicitation was Hardell and Eriksson (1999). From using LPC it emerges that, in this example, less observations are required in the group where prior information is sharper. This results in a more efficient allocation rule than other procedures commonly used in the literature.

Riferimenti bibliografici

- Adcock J. (1997) Sample size determination: a review, *The Statistician*, 46, 261–283.
- Chaloner K. and Verdinelli I. (1995) Bayesian experimental design: a review, *Statistical Science*, 10, 237–308.
- Freireich E. *et al.* (1963) The effect of 6-Mercaptopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy, *Blood*, 21, 699–716.
- Hardell L. and Eriksson M. (1999) A case-control study of a non-Hodgkin lymphoma and exposure to pesticides, *Cancer*, 85, 1353–1360.
- Joseph L. and Belisle P. (1997) Bayesian sample size determination for normal means and difference between normal means, *The Statistician*, 46, 209–226.
- Joseph L., du Berger R. and Belisle P. (1997) Bayesian and mixed Bayesian/likelihood criteria for sample size determination, *Statistics in Medicine*, 16, 769–781.
- Lindley D. (1997) The choice of sample size, *The Statistician*, 46, 129–138.
- Piccinato L. (1996) *Metodi per le decisioni statistiche*, Springer-Verlag Italia.
- Raiffa H. and Schlaifer R. (1961) *Applied statistical decision theory*, Harvard University Graduate School of Business Administration, Boston.