

Semi Automatic Bayesian Sample Size Determination

Scelta della Numerosità Campionaria in Ottica Bayesiana Parzialmente Automatica

Fulvio De Santis

Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università ‘La Sapienza’, P.le A. Moro, 5, 00185, Roma
fulvio.desantis@uniroma1.it

Riassunto: I metodi bayesiani per la scelta della numerosità campionaria si basano sul calcolo di valori attesi o di probabilità rispetto alla distribuzione marginale dei dati, per la cui esistenza è richiesto, in genere, che la distribuzione iniziale del parametro incognito sia propria. In questo lavoro si illustra una metodologia per l’identificazione di distribuzioni a priori definite attraverso l’uso di dati storici sul fenomeno in esame, che permette il controllo del peso di tale informazione nell’analisi.

Keywords: experimental design, objective Bayesian analysis, power priors, sample size, statistical evidence.

1. Introduction

Suppose we are interested in choosing the size n of a random sample $\mathbf{X}_n$ whose joint density function, denoted as $f_n(\cdot; \theta)$, depends on an unknown parameter vector θ . The sample size determination (SSD) criterion to use is related to the inferential scope of the experiment. Typically, one is interested in using the data to estimate θ and one wants such inference to be as precise as possible. Adopting the Bayesian approach, given the data $\mathbf{x}_n = (x_1, \dots, x_n)$, the likelihood function $L(\theta; \mathbf{x}_n) = f_n(\mathbf{x}_n; \theta)$ and a prior distribution for θ , $\pi(\cdot)$, inference is based on elaborations of the posterior distribution: $\pi(\theta|\mathbf{x}_n) = f_n(\mathbf{x}_n; \theta)\pi(\theta)/m_n(\mathbf{x}_n)$, where $m_n(\mathbf{x}_n) = \int_{\Theta} f_n(\mathbf{x}_n; \theta)\pi(\theta)d\theta$ is the marginal distribution of the data. Specifically, suppose we are interested in interval estimation of a continuous scalar parameter θ with unimodal posterior density, and that we want to select n so that the observed length of the interval, $\ell(\mathbf{x}_n)$, is “possibly” small. Before observing the data $\mathbf{x}_n$, $\ell(\mathbf{X}_n)$ is a random variable. The idea is to select n so that the chances are high that the observed value $\ell(\mathbf{x}_n)$ of $\ell(\mathbf{X}_n)$ is small, providing accurate information on θ . Pre-experimental computations needed for SSD are made with respect to m_n . Even though the SSD problem can be addressed using several alternative criteria, in the following we focus on the so called *Average Length Criterion* (see Joseph *et al.* (1997) for details): one selects the smallest n such that the expected length of a posterior set has length less than a chosen threshold, the expectation being with respect to the marginal density m_n . Note that, in order to implement the above criterion, we make use of the prior distribution of θ both in the pre-experimental phase, to define m_n , and in the posterior analysis, to obtain $\pi(\theta|\mathbf{x}_n)$ and $\ell(\mathbf{x}_n)$. This is a typical aspect of Bayesian SSD methods. In this paper interest focuses on developing a noninformative approach to the SSD problem. However, noninformative priors are typically improper. While the posterior distribution is, at least in standard problems, proper, the use of an improper prior implies that the marginal is not defined. The

way out is then to consider two distinct priors: a proper *design prior*, to be used to obtain m_n ; and an *analysis prior*, to be used to obtain the posterior. A typical situation in which this kind of analysis might be of interest is when one decides to use an interval estimate of θ that is acceptable both for Bayesian and frequentist statisticians, as it is often the case using Bayesian intervals derived from a noninformative analysis. However, at the same time, one might want to use prior information in the design stage, in order to take into account uncertainty on the unknown parameter and, also, to avoid local optimality. This is, for instance, the approach followed by Joseph *et al.* (1997) and Wang and Gelfand (2002). Here we suggest essentially the same approach as Joseph *et al.* (1997) in using noninformative priors in computing posterior quantities, but we also propose a method to control the impact of the prior in the marginal and to module the role pre-experimental information has in the SSD process. The basic idea is that previous studies may not be directly related to the current experiment, motivating the need of down-weighting such sources of external evidence. The problem of “discounting” pre-experimental evidence in Bayesian analysis has been considered, among others, by Ibrahim and Chen (2000), who propose the so-called class of power priors. The specific contributions of the present paper are: *i*) the proposal of a semi-automatic and constructive method to derive a class of priors with the above-mentioned discounting property that, in some cases but not in general, coincide with power priors; *ii*) the use of these priors in sample size determination problems.

2. Geometric Priors

Suppose we have historical data, $\mathbf{z}_{n_0}$ of size n_0 , to be used to elicit a design prior. As stated above, lack of homogeneity of historical information with the data to be observed might reasonably suggest to “discount” the influence of $\mathbf{z}_{n_0}$ in the analysis. In order to construct a proper historical-data-dependent design prior a natural approach is to start with a noninformative prior, $\pi^N(\theta)$, and to consider a subset of $\mathbf{z}_{n_0}$, $\mathbf{z}_{n_0}(r)$, of size r , $1 \leq r \leq n_0$, as a *training sample* to convert the starting prior into a proper posterior: $\pi(\theta|\mathbf{z}_{n_0}(r)) \propto \pi^N(\theta) \times L(\theta; \mathbf{z}_{n_0}(r))$. A training sample $\mathbf{z}_{n_0}(r)$ is *proper* if $\pi(\theta|\mathbf{z}_{n_0}(r))$ is proper and is *minimal* if it is proper and none of its subsets is proper. Of course, the smaller the size r of the training sample, the less informative the corresponding posterior that, nevertheless, depends on the specific selected subset $\mathbf{z}_{n_0}(r)$ of $\mathbf{z}_{n_0}$. One can eliminate such dependence by replacing the likelihood function $L(\theta; \mathbf{z}_{n_0}(r))$ with an average likelihood obtained by considering all possible training samples of a certain size: $\text{Average}_{\mathbf{z}_{n_0}(r) \in \mathcal{Z}_r} [L(\theta; \mathbf{z}_{n_0}(r))]$, where $\mathcal{Z}_r$ is the set of all proper training samples of $\mathbf{z}_{n_0}$ of size r . One standard way to proceed is to perform the geometric mean of the likelihoods. The result of this procedure, originally proposed by De Santis and Spezzaferrri (2001), is the *geometric likelihood*:

$$L^G(\theta; \mathbf{z}_{n_0}, r) \propto \left[\prod_{\mathbf{z}_{n_0}(r) \in \mathcal{Z}_r} L(\theta; \mathbf{z}_{n_0}(r)) \right]^{\frac{1}{T}},$$

where T is the number of training samples of size r in $\mathcal{Z}_r$. Intuitively, L^G provides a sort of “average” information from $\mathbf{z}_{n_0}$, with the strength of a sample of size r . Denoting as $a_0 = r/n_0$ the ratio between the sizes of the training samples and of the historical data set, the *geometric prior* is finally defined as the product of the starting prior times the

geometric likelihood:

$$\pi^G(\theta|\mathbf{z}_{n_0}, a_0) \propto \pi^N(\theta) \times L^G(\theta; \mathbf{z}_{n_0}, a_0 n_0).$$

For i.i.d. data, it is straightforward to show that the geometric likelihood is proportional to likelihood for the historical data, $\mathbf{z}_{n_0}$, raised to the fraction $a_0 = r/n_0$: $\pi^G(\theta|\mathbf{z}_{n_0}, a_0) \propto \pi^N(\theta) \times \{L(\theta; \mathbf{z}_{n_0})\}^{a_0}$. Hence, r is the *weight* of the historical data one wants to take into account in the design stage. Note that this is a special case of the power prior by Ibrahim and Chen (2000). Derivation of the geometric prior is, often, straightforward. This is the case with models that belong to the exponential family. For instance, it is easy to check that, assuming a $N(\theta, 1)$ model, given the historical data $\mathbf{z}_{n_0}$ with mean $\bar{z}_{n_0}$, the geometric prior for θ is $N(\bar{z}_{n_0}, [a_0 n_0]^{-1})$. The fraction a_0 has, in the prior, the role of a scale parameter that “discounts” the strength n_0 of the evidence from historical data. Properties of geometric priors for exponential families, extensions to the case in which the data $\mathbf{z}_{n_0}$ are not exchangeable and computational issues for more complex problems are considered in De Santis (2003).

3. Sample Size Determination for Inference on a Proportion

In this section we consider the sample size determination problem for the parameter of a Bernoulli distribution, using the Average Length Criterion. This problem has been previously considered by other authors, using several criteria. Assuming exchangeability of the data, we have that the likelihood function is: $L(\theta; \mathbf{x}_n) = \theta^s (1 - \theta)^{n-s}$ where $s = \sum_{i=1}^n x_i$ denotes the number of “successes” in n Bernoulli trials. Let us consider a uniform noninformative analysis prior for θ . Then, the posterior density is $\text{Beta}(s + 1, 1 + n - s)$. Using the standard non informative prior $\pi^N(\theta) = 1$, the design geometric prior turns out to be $\text{Beta}(a_0 s_0 + 1, a_0(n_0 - s_0) + 1)$, where s_0 denotes the number of successes in the n_0 historical data, $\mathbf{z}_{n_0}$. From standard conjugate analysis it follows that the marginal distribution of the sufficient statistics, s , needed for SSD computations, is binomial-beta with parameters $(a_0 s_0 + 1, a_0(n_0 - s_0) + 1, n)$.

EXAMPLE. Let us consider an example by Greenhouse and Wasserman (1995), who describe a trial designed to demonstrate whether the joint administration of a biological agent, granulocyte colony-stimulating factor (G-CSF) with a standard chemotherapy, 5-fluorouracil and low-dose leucovorin (5-FU/LV), would be able to lessen the incidence of mucositis. In the study, 35 patients were enrolled but, as a matter of fact, the study was stopped after that 3 of the 4 first patients of the study developed a life-threatening disease (grade 4 leukopenia). Let θ denote the incidence rate of grade 4 leukopenia in a trial, that, following Greenhouse and Wasserman (1995), we refer to as toxicity rate, and suppose we are interested in estimating such a quantity using, for instance, a credible set. In a previous randomized trial, 12 out of 176 patients who received a similar treatment (same dose of chemotherapy but without the biological agent) developed grade 4 leukopenia. As in Greenhouse and Wasserman (1995) we want to use this information in order to specify a prior distribution for θ . Since the earlier study was historical and included previously untreated patients with different cancers than the current study, data from the earlier and the new study cannot be considered exchangeable. This motivates the need of down-weighting the information from the previous study. Then, let us set $n_0 = 176$, $s_0 = 12$ and consider several values of the penalizing coefficient, a_0 . It can be checked

that the stronger the weight given to prior information, the smaller the expected length of the posterior interval. For instance, if we give full weight ($a_0 = 1$) to prior information, the average length of the 0.95 equal-tails intervals with $n = 35$ observations in the new study is 0.173, while assuming, $a_0 = 8/n_0$ (corresponding to assuming a “prior sample size” of 8 observations) the average length is 0.215. As an extreme choice, assuming $a_0 = 1/n_0$ (design prior weight of 1 observations, that, in this example, is the minimal training sample), the expected length of the .95 equal-tails interval becomes 0.256. Fig. 1 shows, for illustration, the variation of the expected length of the equal-tails intervals as functions of n , for the three values of a_0 used above: the stronger the weight of prior information (the larger a_0), the smaller the corresponding minimal sample size such that the expected length of the interval is less than the chosen value of the threshold.

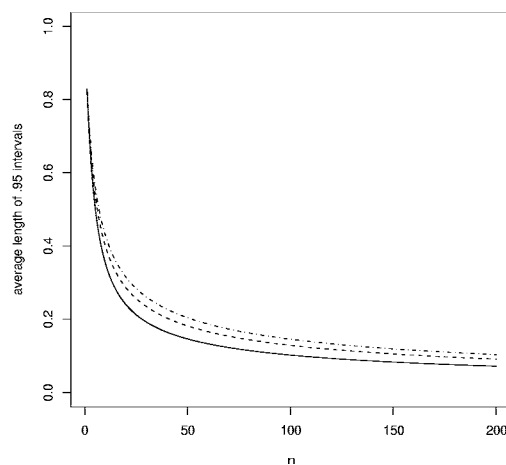


Figure 1: Average length of .95 equal-tails intervals of the Bernoulli parameters as a function of n when $a_0 = 1$ (solid line), $a_0 = 8/n_0$ (dashed line), $a_0 = 3/n_0$ (dashed-dotted line) and with $n_0 = 176$ and $s_0 = 12$.

References

- De Santis F. (2003) A class of priors for semi automatic Bayesian sample size determination, *Rapporto Tecnico n. 17, Dip.to di Statistica, Probabilità e Statistiche Applicate. Università “La Sapienza”*.
- De Santis F. and Spezzaferri F. (2001) Consistent fractional Bayes factor for nested normal linear models, *Journal of Statistical Planning and Inference*, 97, 305–321.
- Greenhouse J. and Wasserman L. (1995) Robust Bayesian methods for monitoring clinical trials, *Statistics in Medicine*, 1, 1379–1391.
- Ibrahim J. and Chen M. (2000) Power prior distributions for regression models, *Statistical Science*, 15, 46–60.
- 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.
- Wang F. and Gelfand A. (2002) A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models, *Statistical Science*, 17, 193–208.