J. R. Statist. Soc. A (2007)
170, *Part* 1, *pp.* 95–113

Using historical data for Bayesian sample size determination

Fulvio De Santis

Università di Roma "La Sapienza", Italy

[Received July 2004. Final revision April 2006]

Summary. We consider the sample size determination (SSD) problem, which is a basic yet extremely important aspect of experimental design. Specifically, we deal with the Bayesian approach to SSD, which gives researchers the possibility of taking into account pre-experimental information and uncertainty on unknown parameters. At the design stage, this fact offers the advantage of removing or mitigating typical drawbacks of classical methods, which might lead to serious miscalculation of the sample size. In this context, the leading idea is to choose the minimal sample size that guarantees a probabilistic control on the performance of quantities that are derived from the posterior distribution and used for inference on parameters of interest. We are concerned with the use of historical data—i.e. observations from previous similar studies—for SSD. We illustrate how the class of power priors can be fruitfully employed to deal with lack of homogeneity between historical data and observations of the upcoming experiment. This problem, in fact, determines the necessity of discounting prior information and of evaluating the effect of heterogeneity on the optimal sample size. Some of the most popular Bayesian SSD methods are reviewed and their use, in concert with power priors, is illustrated in several medical experimental contexts.

Keywords: Elicitation; Experimental design; Historical data; Power priors; Sample size; Statistical evidence

1. Introduction

The basic sample size determination (SSD) problem can be introduced by a simple example. Suppose that we want to estimate an unknown quantity θ that is related to a population of interest. Without loss of generality, let us confine ourselves to the medical experimental context. In this case, θ may represent, for instance, the effect of a new drug, the expected survival time due to an innovative treatment or the proportion of patients in the population who respond to a therapy. In the planning stage of the experiment aimed at estimating θ, we want to select a number of observations that is sufficiently large to guarantee good quality inference. Depending on the specific inferential goal that we have (estimation or testing, for instance) and on the inferential approach that we follow (frequentist or Bayesian, for instance), a considerable number of alternative SSD criteria are available. This is true especially from a frequentist perspective. The most common classical criteria are based on the idea of controlling some aspects of the sampling distributions of statistics that are used for inference. In estimation problems, for instance, we aim to control either the size of interval or the variance of point estimators. In this regard, computations for SSD are performed under the sampling distribution of the data. Hence, resulting criteria typically depend on one or more unknown parameters, and initial

Address for correspondence: Fulvio De Santis, Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università di Roma "La Sapienza", Piazzale A. Moro 5, 00185 Roma, Italy.
E-mail: fulvio.desantis@uniroma1.it

 0964–1998/07/170095

guesses of the true values of the parameters are needed for implementation of these procedures. For example, sample size formulae for testing a normal mean depend on the variance that must be replaced by guessed values. Often, as in the well-known problem of choosing the sample size for interval estimation of a binomial proportion, classical procedures depend directly on the unknown parameter of interest. Therefore, the resulting sample sizes are only locally optimal and can depend quite dramatically on the chosen design value. For an overview of frequentist SSD, see, for instance, Desu and Raghavarao (1990). See also Julious (2004) for applications to clinical trials.

The local optimality problem of the frequentist approach is not shared by Bayesian methods, which allow statisticians to model uncertainty on both interest and nuisance parameters via prior distributions. More specifically, Bayesian inferential methods (see, for instance Spiegelhalter *et al*. (2004) and O'Hagan and Forster (2004)) are based on elaborations of the posterior distribution, which synthesize pre-experimental information on θ (prior distribution), and experimental data (likelihood). In this context, an adequate sample size is chosen to control probabilistically the performance of certain aspects of the posterior distribution, such as the precision of posterior point or interval estimates. The posterior distribution and its functionals depend on the data that, before being observed, are random. The probability distribution of the data that is used for Bayesian SSD is the marginal (or prior predictive) distribution, i.e. a mixture of the sampling distribution of the data with respect to the prior distribution for unknown parameters. Consequently, the resulting sample sizes do not depend on specific guessed values for the unknown parameters but rather on their prior distributions. Furthermore, the use of this prior information might help to avoid basing the design of experiment on, for instance, overenthusiastic beliefs on θ, with the potential consequence of serious miscalculations of the sample size (Spiegelhalter *et al*. (2004), section 6.5, and Fayers *et al*. (2000)). All this results in greater flexibility than the classical approach even though priors depend in general on hyperparameters that must be specified to implement the analysis.

The literature on Bayesian SSD has recently received considerable attention. There are, for instance, two special issues of *The Statistician* with several contributions to the topic (see volume 44, part 2 (1995), and volume 46, part 2 (1997)). Among others, see Adcock (1997) and Joseph and Belisle (1997). For more recent contributions, see Wang and Gelfand (2002), De Santis and Perone Pacifico (2003), Clarke and Yuan (2006) and De Santis (2006a). For the important part of the literature on Bayesian SSD, which approaches the problem from a decision theoretical point of view, see the pioneering work of Raiffa and Schlaifer (1961) and, more recently, Bernardo (1997), Lindley (1997) and Walker (2003). For excellent reviews on the more general topic of Bayesian experimental design, see Chaloner and Verdinelli (1995) and DasGupta (1996).

A specific feature of Bayesian SSD criteria is that their implementation requires the prior distribution to be used twice: both to obtain the posterior for θ and also to define the prior predictive distribution for preposterior computations. A large part of the literature has developed criteria that employ the same (proper) prior both at the design and at the posterior stage. Nevertheless, several researchers, like Joseph and Belisle (1997) and Spiegelhalter *et al*. (2004), have pointed out that there are contexts in which, even in the presence of substantial prior knowledge on the unknown parameter, this information cannot be included in final inference on θ; this is often required, for instance, by regulatory authorities in medical studies. A way out is to exploit pre-experimental information for elicitation of a prior distribution to be used at the design stage and to use a non-informative prior for obtaining a posterior distribution for final analysis. This corresponds essentially to a hybrid frequentist–Bayesian approach 'in which prior information is formally used but final analysis is carried out in a classical framework' (Spiegelhalter *et al*. (2004), section 6.5). This mixed approach to SSD has been already proposed, for instance, by

Spiegelhalter and Freedman (1986) and by Joseph *et al*. (1997). A more general two-priors procedure, which is not necessarily based on non-informative distributions for final analysis, has recently been motivated and discussed by Wang and Gelfand (2002). See also Sahu and Smith (2004) and De Santis (2006a) for related ideas.

This paper deals with the use of *historical data*, i.e. results from previous similar experiments on θ, in Bayesian SSD. Specifically, we consider the two-priors approach in which historical information is employed solely for the design of the trial whereas standard non-informative priors are used to obtain posterior quantities that are not affected by extra-experimental knowledge and that can be acceptable also from a frequentist perspective. The main goal of the paper is to suggest a procedure for building up a design prior based on available historical data. We propose to employ for design the class of power priors, which was introduced by Ibrahim and Chen (2000) in the context of posterior analysis. The method is easy to implement and allows us to control the effect of prior information in the SSD process. This latter aspect is relevant for two reasons. The first is that data from previous and future studies are not necessarily homogeneous, and we might want to take this into account by discounting historical evidence. The second reason is that, if the size of historical data is too large, the resulting design prior might fail to account properly for uncertainty on the value of the parameter in planning the new experiment.

The paper is organized as follows. Section 2 introduces and formalizes the Bayesian approach to SSD. Section 2.1 presents a review of some Bayesian SSD criteria that have been proposed in the literature, whereas Section 2.2 focuses on the distinction between priors for design and priors for posterior analysis. Section 3 is about the use of historical data for the construction of a prior for the design of trials: the power prior method is here described and discussed. Section 4 deals with computational issues that are related to implementation of the SSD methods under consideration and Section 5 illustrates some examples. Section 6 presents a generalization of the method and, finally, Section 7 contains a discussion.

2. Bayesian sample size determination

Suppose that we are interested in choosing the size n of a random sample $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ whose joint density function $f_n(\cdot|\theta)$ (we use here the notation for continuous random variables, without loss of generality) depends on an unknown parameter vector θ that we want to estimate. Adopting the Bayesian approach, given the data $\mathbf{x}_n = (x_1, \ldots, x_n)$, the likelihood function $L(\theta; \mathbf{x}_n) \propto f_n(\mathbf{x}_n|\theta)$ and the prior distribution $\pi(\cdot)$ for the parameter, inference is based on elaborations of the posterior distribution of θ:

$$\pi(\theta|\mathbf{x}_n) = \frac{f_n(\mathbf{x}_n|\theta)\,\pi(\theta)}{\int_\Theta f_n(\mathbf{x}_n|\theta)\,\pi(\theta)\,\mathrm{d}\theta}.$$

Let $T(\mathbf{x}_n)$ denote a generic functional of the posterior distribution of θ whose performance we want to control by designing the experiment. For instance, $T(\mathbf{x}_n)$ might be either the posterior variance, or the width of the highest posterior density (HPD) set or the posterior probability of a certain hypothesis. Before observing the data $\mathbf{x}_n$, $T(\mathbf{X}_n)$ is a random variable. The idea is to select n so that the observed value $T(\mathbf{x}_n)$ of $T(\mathbf{X}_n)$ is likely to provide accurate information on θ. Pre-experimental computations for SSD are made with the marginal density function of the data,

$$m_n(\mathbf{x}_n; \pi) = \int_\theta f_n(\mathbf{x}_n|\theta)\,\pi(\theta)\,\mathrm{d}\theta,$$

a mixture of the sampling distribution and the prior distribution of θ. In what follows, we shall denote with $\mathbb{P}$ the probability measure corresponding to the density m_n and with $\mathbb{E}$ the expected

value computed with respect to m_n. Most Bayesian SSD criteria select the minimal n so that, for chosen values $\varepsilon > 0$ and $\varepsilon' \in (0, 1)$, one of the two following statements is satisfied:

$$\mathbb{E}[T(\mathbf{X}_n)] \leqslant \varepsilon \tag{1}$$

or

$$\mathbb{P}\{T(\mathbf{X}_n) \in \mathcal{A}\} \leqslant \varepsilon', \tag{2}$$

where $\mathcal{A}$ is a subset of the values that the random variable $T(\mathbf{X}_n)$ can assume. Therefore, the SSD determination problem and the following inference is a two-step process: first select n^* by a preposterior calculation; then, use $\pi(\theta|\mathbf{x}_{n^*})$ to obtain $T(\mathbf{x}_{n^*})$.

2.1. Sample size determination criteria

Let us now review some Bayesian SSD criteria for estimation, assuming for simplicity that θ is a scalar parameter. See, for instance, Wang and Gelfand (2002) for extensions to the multi-parameter case. The following are three examples of criteria that control the performance of standard estimation methods. Criteria (b) and (c) are interval-type methods, based on the idea of controlling aspects of the probability distribution of the random length of credible sets.

(a) In the average posterior variance criterion, for a given $\varepsilon > 0$, choose the smallest n such that

$$\mathbb{E}[\mathrm{var}(\theta|\mathbf{X}_n)] \leqslant \varepsilon,$$

where $\mathrm{var}(\theta|\mathbf{X}_n)$ is the posterior variance of θ. In this case, $T(\mathbf{x}_n) = \mathrm{var}(\theta|\mathbf{x}_n)$. This criterion controls the dispersion of the posterior distribution. Of course, we can decide to use different dispersion measures, deriving alternative criteria.

(b) In the average length criterion (ALC), for a given $l > 0$, we look for the smallest n such that

$$\mathbb{E}[L_\alpha(\mathbf{X}_n)] \leqslant l, \tag{3}$$

where, for a fixed $\alpha \in (0, 1)$, $L_\alpha(\mathbf{X}_n)$ is the random length of the $(1-\alpha)$-level posterior set for θ. Here, $T(\mathbf{x}_n) = L_\alpha(\mathbf{x}_n)$. In what follows we shall limit attention either to HPD sets (i.e. subsets of the parameter space whose points have posterior density that is not smaller than a given level) or to equal-tails sets. Note that this criterion, which was proposed by Joseph *et al.* (1995), controls the average length of the HPD set, but not its variability.

(c) In the length probability criterion (LPC), for given $l > 0$ and $\varepsilon' \in (0, 1)$, choose the smallest n such that

$$\mathbb{P}\{L_\alpha(\mathbf{X}_n) \geqslant l\} \leqslant \varepsilon'. \tag{4}$$

As for the ALC, $T(\mathbf{x}_n) = L_\alpha(\mathbf{X}_n)$ and the LPC can be written in the general form (2), with $\mathcal{A} = (l, U)$, where U denotes the upper bound for the length of the credible interval. Joseph and Belisle (1997) derived the LPC as a special case of the *worst outcome criterion*, which was originally introduced by Joseph *et al.* (1995). See also Joseph *et al.* (1997) and De Santis and Perone Pacifico (2003) for further details.

Explicit expressions for these SSD estimation criteria can be obtained only in very specific cases. In general, we must resort to numerical approximations. These issues are discussed in Section 4. Numerical applications and examples are considered in Section 5.

SSD criteria for model selection and hypothesis testing are also available in the literature but, for brevity, they are just mentioned here. The idea is that, in the presence of several alternative

models and for a given choice criterion, we search for the minimal number of observations for which the correct model is likely to be selected. For details on this topic, see, among others, Weiss (1997), Wang and Gelfand (2002) and De Santis (2004).

2.2. The two-priors approach

As stated in Section 1, we here follow the two-priors approach to SSD, which makes a sharp distinction between the *analysis prior*, which is used to compute the posterior and to determine $T(\mathbf{x}_n)$, and the *design prior*, which is used to obtain m_n. The motivation for adopting separate priors for θ, which was thoroughly discussed by Wang and Gelfand (2002) and Sahu and Smith (2004), is that they play two distinct roles in design and inference steps. The analysis prior formalizes pre-trial knowledge that we want to take into account, together with experimental evidence, in the final analysis. The design prior describes a scenario—which is not necessarily coincident with the description of the analysis prior—under which we want to select the sample size. It serves to obtain a marginal distribution m_n that incorporates uncertainty on a guessed value for θ. The design prior is hence used, according to Wang and Gelfand (2002), in a 'what if?' spirit: assuming that θ is highly likely to be in a certain subset of the parameter space, and assuming model uncertainty on θ according to a specific distribution, what are the consequences in terms of the predictive distribution of a posterior quantity of interest?

2.2.1. Example

Suppose that θ represents the difference between the effects of a new and a standard therapy, positive values implying superiority of the new treatment. Suppose also that a former study has yielded an estimate θ_0 for θ, which provided evidence in favour of the new treatment. We now want to plan a new experiment to confirm results of the former study. In this case, it is natural to assume, as the design prior, a distribution that is centred on θ_0. However, if we want to report final inference which reflects scientific pre-experimental neutrality between the two alternative candidate therapies, as is typically requested by regulatory authorities, the posterior distribution of θ is to be defined by using a prior that is centred on zero. Hence, in this case, the analysis and design priors must differ, at least, for the location parameter.

As mentioned in Section 1, in this paper we consider Bayesian point or interval estimates that are derived by using non-informative analysis priors. In many circumstances, these procedures coincide with standard classical estimation tools and are therefore acceptable also from a frequentist viewpoint. At the same time, we exploit prior information in the design stage, to take into account uncertainty on unknown parameters, to avoid local optimality and miscalculations of the sample size. The availability of substantial prior knowledge for the design of the trial, which is assumed in this paper and that results in a proper design prior, is also relevant from a technical point of view. In fact, non-informative priors may often be improper and cannot be used for design, since the resulting marginal distribution may not even exist, the integral that defines m_n being divergent. Therefore, in general, unlike the analysis prior that can be improper as long as the resulting posterior is proper, the analysis prior must be proper.

3. Power priors for design

In this section we review the class of power priors (see Ibrahim and Chen (2000) and Ibrahim *et al*. (2001)) as a method for incorporating historical information in a design prior and also for deciding the weight that such information has in the SSD process.

Let $\mathbf{z}_{n_0}$ be a sample of size n_0 of historical data from a previous study, $L(\theta; \mathbf{z}_{n_0})$ the likelihood function of θ based on these data and $\pi_0(\theta)$ a prior that we would use for inference on θ if we had no further information. The use of the same likelihood function $L(\theta; \cdot)$ for data from the previous and the future experiments is an implicit assumption of homogeneity between $\mathbf{z}_{n_0}$ and $\mathbf{x}_n$. Power priors are defined hierarchically as follows. Consider the posterior $\pi^{\mathrm{P}}(\cdot|\mathbf{z}_{n_0}, a_0)$, which is obtained by combining the prior $\pi_0(\cdot)$ and the likelihood $L(\theta; \mathbf{z}_{n_0})$, suitably scaled by an exponential factor a_0:

$$\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0) \propto \pi_0(\theta)\, L(\theta; \mathbf{z}_{n_0})^{a_0}, \qquad a_0 \in (0, 1). \tag{5}$$

If π_0 is proper, $\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)$ is also proper; if π_0 is improper, $\mathbf{z}_{n_0}$ must be such that $\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)$ is proper. The coefficient a_0 has the effect of measuring out the importance of historical data in $\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)$. As $a_0 \to 1$, we obtain the standard posterior of θ given $\mathbf{z}_{n_0}$; as $a_0 \to 0$, $\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)$ tends to the initial prior, π_0; intermediate choices of a_0 result in alternative weights assigned to the information that is conveyed by $\mathbf{z}_{n_0}$ in the posterior. Therefore, a_0 controls the influence that $\mathbf{z}_{n_0}$ has on the analysis and we can consider different discounts of historical evidence by assigning small values to this parameter. See also Spiegelhalter *et al.* (2004), section 5.4, for discussion. The definition of power priors is completed by considering a mixture of the above priors with respect to a mixing distribution for a_0. The effect of mixing is to obtain a prior for θ that has, in general, heavier tails than those which are obtained for fixed a_0. Of, course, elicitation of a prior for the weight parameter a_0 is a crucial point. However, in what follows we shall consider only the case in which a_0 is not random. See Ibrahim and Chen (2000) for discussion.

Turning to the SSD problem, in which data $\mathbf{X}_n$ are still to be observed, we assume that π_0 is a non-informative prior and we consider $\pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)$ as the design prior to define a proper marginal distribution of the data $\mathbf{X}_n$:

$$m_n(\mathbf{x}_n|\mathbf{z}_{n_0}, a_0) = \int_\Theta f_n(\mathbf{x}_n|\theta)\, \pi^{\mathrm{P}}(\theta|\mathbf{z}_{n_0}, a_0)\, \mathrm{d}\theta.$$

This distribution and the resulting sample sizes depend on the value of a_0.

The use of power priors for posterior inference has been criticized (see, for instance, Spiegelhalter *et al.* (2004), page 131, for references) as it does not have any operational interpretation and, as a consequence, as it does not provide a way to assess a value for a_0. Two interpretations of the power prior in the set-up of independent and identically distributed (IID) data are discussed in De Santis (2006b). First, when a maximum likelihood estimator for θ exists and is unique, π^{P} is equivalent to a posterior distribution that is obtained by using a sample of size $r = a_0 n_0$, which provides the same maximum likelihood estimator for θ as the entire sample $\mathbf{z}_{n_0}$. Second, when the model $f_n(\cdot|\theta)$ belongs to the exponential family, the prior π^{P} for the natural parameter coincides with the standard posterior distribution that is obtained by using a sample of size $r = a_0 n_0$ whose arithmetic mean is equal to the historical data mean. Hence, at least in some standard problems, a power prior can be interpreted as a posterior distribution that is associated with a sample whose informative content on θ is *qualitatively* the same as that of the historical data set, but *quantitatively* equivalent to that of a sample of size r.

The use of fractional likelihoods as a basis for the construction of priors has a well-established tradition in Bayesian statistics. O'Hagan (1995) and De Santis and Spezzaferri (1997, 2001) used fractional likelihoods for defining weakly data-dependent priors in Bayesian model selection. Borrowing ideas from these, in De Santis (2006b) it is pointed out that, for IID historical data, the fractional likelihood in expression (5) is the geometric mean of the likelihoods for θ associated with all the possible subsamples of $\mathbf{z}_{n_0}$ of size $r = a_0 n_0$. Hence, the power prior has

the interpretation of a posterior distribution determined with an average likelihood and whose informational strength is that of a sample of size $r < n_0$. Furthermore, the approach that is based on the geometric mean gives a simple algorithm for constructing power priors also for non-IID data. This procedure leads to a generalization of the basic single-fraction form of power priors. See Section 6 for an example.

3.1. Choosing the fraction a_0

The choice of the fraction a_0 is of course crucial in the power prior approach. This is true both when power priors are used for posterior analysis and also, as we shall see, when they are employed as design priors in SSD. In general, there is no formal way for choosing a_0 and the amount of discount of past evidence depends on the trust in historical data and on their homogeneity with data from the new experiment. For instance, in the context of posterior analysis, Greenhouse and Wasserman (1995) downweighted a previous trial with 176 subjects to be equivalent to only 10 patients. In this case the discount was motivated by a lack of homogeneity between patients in the two studies and also by the necessity of avoiding evidence from the past study overwhelming information from the new experiment. More recently, Fryback *et al.* (2001) used power priors for the comparison of new and traditional therapies for myocardial infarction and proposed radical discount of past evidence, to account for substantial changes in the protocol of a future study.

In the SSD context, the choice of a_0 is reflected by the value of the optimal sample size. For the criteria that are used in the following examples (Section 5), for instance, optimal sample sizes are decreasing functions of a_0. This means that, as a_0 (the weight that is assigned to pre-trial information) increases, the corresponding minimal sample size that is required to achieve a prespecified target decreases. Hence, in the examples that are considered, the stronger the confidence that we have in historical data, the larger the number of units we can save in the new experiment. The intuitive explanation for this is that increasing the weight of historical data has the effect of reducing variability of the marginal distribution, i.e. uncertainty associated with the data generator mechanism. In the same examples, it is also shown that the decrease in the optimal sample size as a_0 increases is not linear and that there are situations in which even a small increase in a_0 may determine a dramatic reduction in the corresponding sample size. Hence, even mild confidence in the historical data may sometimes result in a considerable saving in sample size.

From a practical perspective, we can define ranges of values for a_0 corresponding to different degrees of discount of historical evidence. For instance, without any pretence of generality, we can use for reference the classification in Table 1 and choose a_0 accordingly. However, there is unavoidable arbitrariness in the definition of any possible scale of discount levels. Hence, from a pragmatic viewpoint, it seems more appealing to evaluate the effect of the historical information on the optimal sample size by drawing plots of n^* as a function of a_0. By looking at these plots

Table 1. Levels of discount

a_0	*Discount*
<0.2	Severe
0.2–0.5	Substantial
0.5–0.8	Moderate
>0.8	Slight

we can establish, for instance, whether the level of trust in historical data leads to sample sizes that are affordable or compatible with the financial constraints on the research. Or, conversely, we can check whether the number of sample data that we can afford corresponds to a value of a_0 which reflects an acceptable degree of trust in historical data.

4. Computations

Derivation of the power prior is often straightforward. As an example, for an $N(\theta, 1)$ model, the power prior for θ is a normal density centred on the arithmetic mean of the historical data set and variance equal to $(a_0 n_0)^{-1}$. Hence, the fraction a_0 is a coefficient of the scale parameter that discounts the evidential strength n_0 of historical data. However, in more general settings, unless conjugate priors are used, closed form expressions for power priors, for the corresponding marginal distributions and for SSD criteria are not available. In these cases, numerical computations based on simulations might be necessary. This approach is illustrated, for instance, in Wang and Gelfand (2002). See also Clarke and Yuan (2006) for an alternative approach based on higher order asymptotic approximations.

The idea of the simulation-based approach to Bayes SSD is as follows. For several values of n, we draw samples $\tilde{\mathbf{x}}_n$ from the predictive distribution m_n of the data. Then, we compute $T(\tilde{\mathbf{x}}_n)$ and, by repeating this operation a large number of times, we obtain an approximate value for sample size criteria (4) or (5). In this way a plot of the SSD quantity as a function of n is obtained and we can choose the minimal n such that it is less than or equal to a chosen threshold. Using the class of power priors, samples $\tilde{\mathbf{x}}_n$ from the predictive distribution m_n can be drawn as follows:

(a) draw $\tilde{\theta}$ from the power prior $\pi^{\mathrm{P}}(\cdot|\mathbf{z}_{n_0}, a_0)$;
(b) draw a sample $\tilde{\mathbf{x}}_n$ from the sampling distribution $f_n(\cdot|\tilde{\theta})$.

This procedure is repeated N times and the following final steps are then performed:

(c) compute $T(\tilde{\mathbf{x}}_n)$, for each of the N generated samples;
(d) approximate $\mathbb{P}\{T(\mathbf{X}_n) \in \mathcal{A}\}$ with the proportion of the N generated samples for which $T(\tilde{\mathbf{x}}_n)$ belongs to the set $\mathcal{A}$. Similarly, $\mathbb{E}[T(\mathbf{X}_n)]$ is approximated by the arithmetic mean of the N values $T(\tilde{\mathbf{x}}_n)$.

Step (c) might be non trivial for interval-based criteria. In particular this is true when the posterior distribution is not symmetric and HPD intervals are considered for implementing the ALC or LPC. De Santis and Perone Pacifico (2003), assuming unimodality of the posterior distribution, proposed a numerical procedure for approximate determination of HPD sets and for simulation of the distribution of the length $L_\alpha(\mathbf{X}_n)$. This procedure, which can be implemented for the ALC and LPC, requires sampling from the posterior distribution; in many applied problems that can either be done directly or by using standard numerical techniques (Wang and Gelfand, 2002).

5. Examples

In this section we consider some basic examples to illustrate the derivation of power priors and of quantities that are necessary for SSD. We consider both cases in which closed form formulae are available (Sections 5.1 and 5.3) as well as a situation in which simulation is necessary (Section 5.2).

5.1. Sample size for inference on the normal mean

Suppose that $X_1, \ldots, X_n$ are IID, normally distributed with mean μ and precision λ both unknown and assume also that μ is the parameter of interest. The normal distribution model is

widely used in clinical trials and the mean parameter might denote, for instance, the effect of a treatment or the difference in the effects of alternative therapies.

Using the non-informative analysis prior $\pi^{\mathrm{N}}(\mu, \lambda) = \lambda^{-1}$, it follows that the posterior distribution of μ is a Student t-density with $n-1$ degrees of freedom and with location and scale parameters respectively equal to the sample mean $\bar{x}_n$ and

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 / n$$

(see Bernardo and Smith (1994), page 440). Given $\mathbf{z}_{n_0}$, the power prior for μ and λ is easily determined and preposterior computations yield closed form expressions for the estimation criteria that were listed in Section 3.1. Technical details are reported in Appendix A.

Let us now consider a few simple numerical examples to illustrate the effect of historical data on the optimal sample size and the role of the discount parameter a_0. Fig. 1 shows the curves of $\mathbb{E}[\mathrm{var}(\mu|\mathbf{X}_n)]$ as functions of n, using the power prior with $s_0 = 1$ and $n_0 = 20$, for three choices of a_0: $a_0 = 1$, corresponding to the use of the full prior for design, $a_0 = 0.5$ and $a_0 = 0.2$, for moderate and severe discount. Using a threshold value $\varepsilon = 0.05$, the optimal sample size is 26 if $a_0 = 1$, 31 if $a_0 = 0.5$ and 82 if $a_0 = 0.2$. As expected, the weight that is assigned to the prior is quite crucial. Specifically, the optimal sample size increases as a_0 decreases. The effect on the optimal sample size of a_0 is graphically illustrated in Fig. 2, whose upper curve shows changes in optimal sample sizes that are obtained when $n_0 = 20$ as a_0 varies in $(a_0^m, 1)$, where $a_0^m = 3/n_0$ is the minimal value for a_0 (see Appendix A). Note that, for small values of a_0, even a moderate increase in a_0 might determine a dramatic decrease in the optimal sample size. This feature is even more remarkable if larger values of n_0 are considered, as is shown by the lower step curve

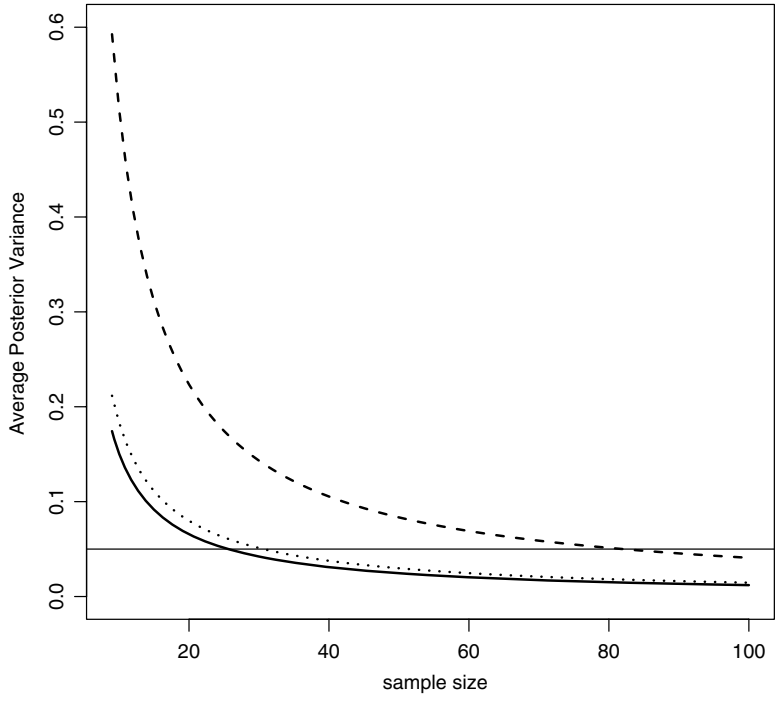

Fig. 1. Average posterior variance criterion for the normal mean, $n_0 = 20$ and $s_0 = 1$, using the power prior with $a_0 = 1$ (———), $a_0 = 0.5$ (· · · · · · ·) and $a_0 = 0.2$ (– – –)

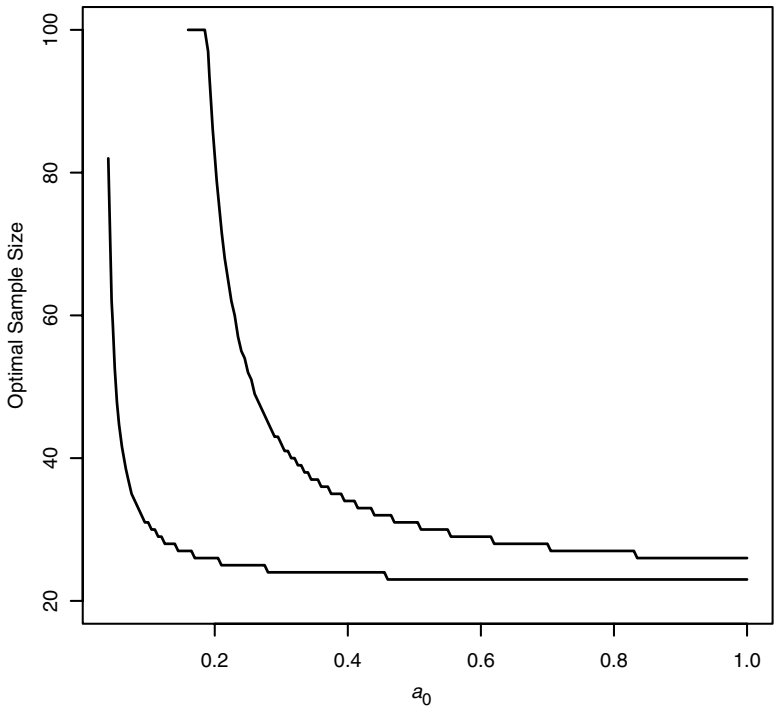

Fig. 2. Optimal sample sizes from the average posterior variance criterion for the normal mean, $s_0 = 1$, using the power prior, as a_0 varies in (0,1), $n_0 = 20$ (upper curve) and $n_0 = 100$ (lower curve)

of Fig. 2 ($n_0 = 100$). In general, the larger n_0 is, the smaller the value of a_0 that is needed to have a substantial reduction in the optimal sample size.

Using the same data ($n_0 = 20$ and $s_0 = 1$), the curves that represent the average length of the 0.95% HPD intervals are plotted, as functions of n, in Fig. 3, for the three values of a_0 that were considered above. Assuming, for example, a threshold level $l = 0.5$, the optimal sample sizes with the ALC are 72, 84 and 158, for a_0 respectively equal to 1, 0.5 and 0.2. As for the average posterior variance criterion example, a 50% discount of historical data has a fairly limited effect on the optimal sample size that is determined with the ALC, whereas only a severe discount, as obtained for $a_0 = 0.2$, determines a considerable increase in the sample size required.

5.2. Sample size for inference on the exponential parameter

Suppose that $X_1, \ldots, X_n$ are exchangeable, exponentially distributed random variables with unknown parameter θ. The exponential model is often used in reliability and survival analysis: in this case the data represent the lifetime of items or subjects and $1/\theta$ represents the expected survival time of the population. A specific feature of reliability and survival analysis is that the data are often censored. Consider for instance type I censoring, and let t_j^* be the censoring time for the jth subject. The likelihood function of θ, for the observed vectors $(\mathbf{t}_n, \boldsymbol{\delta}_n)$, is

$$L(\theta; \mathbf{t}_n, \boldsymbol{\delta}_n) = \prod_{\{j:\,\delta_j=1\}} \theta \exp(-\theta t_j) \prod_{\{j:\,\delta_j=0\}} \exp(-\theta t_j^*)$$
$$= \theta^{\sum_{j=1}^n \delta_j} \exp\left(-\theta \sum_{j=1}^n t_j \right),$$

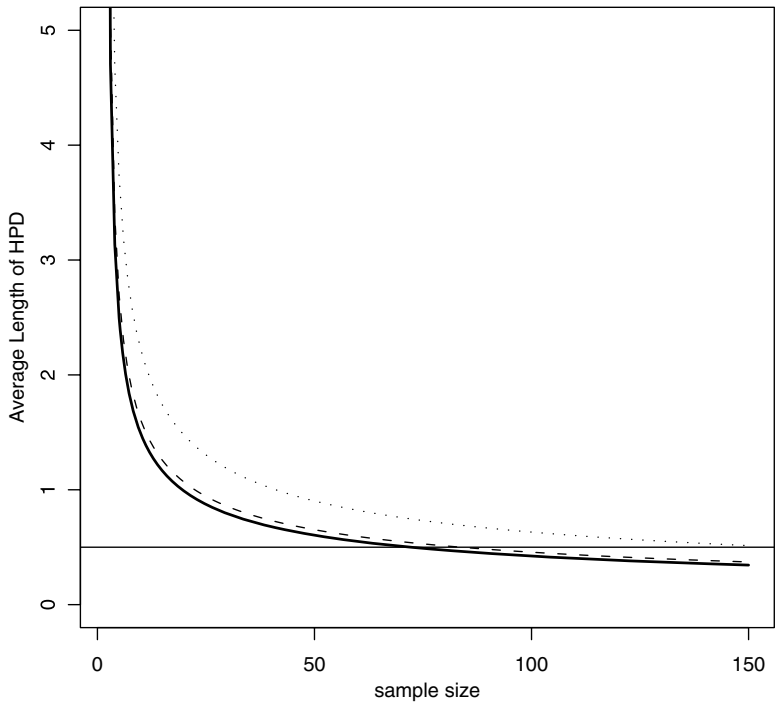

Fig. 3. ALC for the normal mean, $n_0 = 20$, $s_0 = 1$, and $\alpha = 0.05$, using the power prior with $a_0 = 1$ (———), $a_0 = 0.2$ (·····) and $a_0 = 0.5$ (– – –)

where $t_j = \min(x_j, t_j^*)$ is the observed survival time of subject j and $\delta_j = 1$ if $x_j \leqslant t_j^*$ and $\delta_j = 0$ otherwise. Using for θ the standard improper analysis prior $\pi^{\mathrm{N}} \propto \theta^{-1}$, the posterior distribution of θ is a gamma density function of parameters $(\Sigma_{j=1}^{n} \delta_j, \Sigma_{j=1}^{n} t_j)$. Note that asymmetry of this distribution implies that analytic expressions for HPD sets cannot be found and, in the following example, we resort to the numerical procedure of Section 4. Given a set of historical data, it can be checked that the power prior is a gamma density function of parameters $(a_0 \Sigma_{j=1}^{n_0} \delta_j^0, a_0 \Sigma_{j=1}^{n_0} t_j^0)$, where $\Sigma_{j=1}^{n_0} \delta_j^0$ is the total number of uncensored observations and $\Sigma_{j=1}^{n_0} t_j^0$ the sum of survival times in the historical data.

5.2.1. Example

Let us consider SSD for estimation of the mean survival time $1/\theta$ of patients submitted to 6-mercaptopurine therapy, for maintenance of remission in acute leukaemia, a problem which was previously considered in De Santis and Perone Pacifico (2003). Historical data from an old study of Freireich *et al.* (1963) are available: the drug was administered to 21 patients and, at the end of the experiment, the results observed were $\Sigma_{j=1}^{21} t_j = 359$ (in weeks) and $\Sigma_{j=1}^{21} \delta_j = 9$. De Santis and Perone Pacifico (2003) considered a standard Bayesian approach to the SSD problem, using a single proper conjugate prior rather than separate design and analysis priors. They also assumed implicitly exchangeability between data from the old and from the new experiment. However, this is a typical situation in which we might want to use historical information but also to discount its influence. Hence, let us consider, as usual in this paper, a non-informative analysis prior and exploit historical data (nine uncensored observations and total survival 359 weeks) for the trial design. Suppose that we are interested in determining the sample size for an experiment which gives guarantees of finding an interval estimate for $1/\theta$ that is not larger than

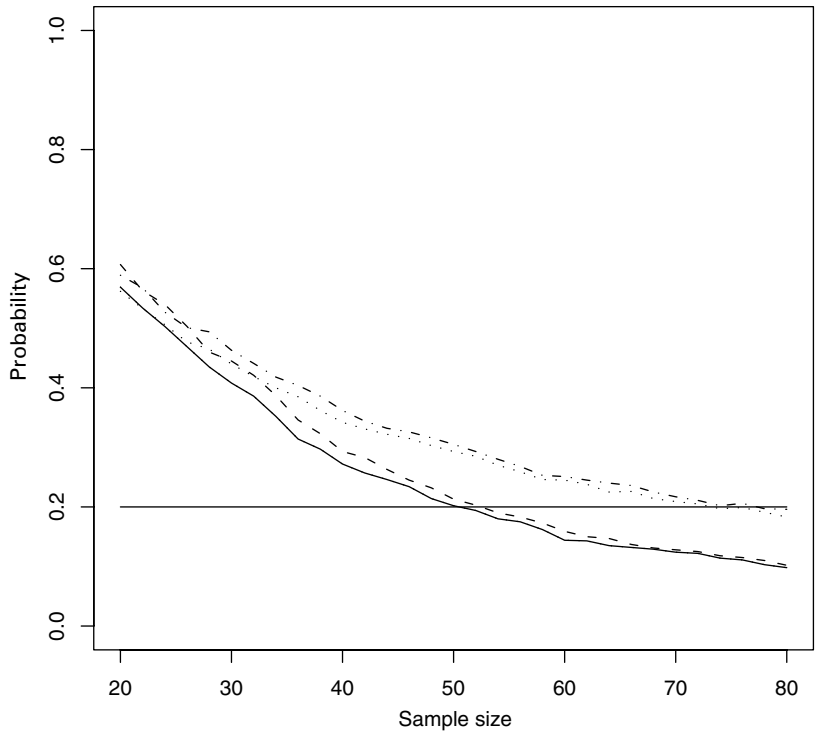

Fig. 4. Survival analysis example: $\Pr\{L_\alpha(\mathbf{X}_n) \geqslant 30\}$ as a function of n when $a_0 = 1$ (———, HPD; – – –, equal tails) and when $a_0 = 0.5$ (·····, HPD; ·–·–·, equal tails)

30 weeks, assuming a censoring time which implies 25% of censored observations in the new experiment. As an example, we compare sample sizes determined by using the LPC, for two values of the discount parameter: $a_0 = 1$ (full weight given to the historical data) and $a_0 = 0.5$ (50% discount of prior knowledge). Fig. 4 shows the plots of $\Pr\{L_\alpha(\mathbf{X}_n) \geqslant 30\}$ as a function of n when $a_0 = 1$ (HPD, full curve; equal tails, broken curve) and when $a_0 = 0.5$ (HPD, dotted curve; equal tails, chain curve). As expected, for a fixed value of a_0 and for each n, the probability that the length of HPD sets is larger than $l = 30$ is smaller than the corresponding probability for equal tails intervals. Also the effect of discount ($a_0 = 0.5$) is relevant here: for instance, considering a threshold $\varepsilon' = 0.2$ for the LPC, the optimal sample size changes from 52 ($a_0 = 1$) to 74 ($a_0 = 0.5$). For equal-tails intervals, the optimal sample size changes from 54 to 78. Finally, using the same prior for the design and analysis as in De Santis and Perone Pacifico (2003), the optimal sample sizes are respectively 26 (HPD) and 29 (equal tails). This shows that taking into account prior information or not for the analysis prior may be strongly influential in the optimal sample sizes that are determined.

5.3. Sample size for inference on a proportion

In this section we consider SSD for estimation of the parameter of a Bernoulli distribution, which represents the proportion of units presenting a characteristic of interest in a population as, for example, the expected fraction of subjects who respond to a certain medical treatment. This problem has been previously considered, for instance, by Joseph *et al.* (1995) and Pham-Gia and Turkkan (1992). Assuming exchangeability of the data, 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. With the

non-informative analysis prior, $\pi^{N}(\theta) = 1$, $\theta \in (0, 1)$, the posterior is a beta density of parameters $(s+1, n-s+1)$. Setting $\pi_0(\theta) = 1$, the design power prior turns out to be a beta density of parameters $(a_0 s_0 + 1, a_0(n_0 - s_0) + 1)$, where s_0 denotes the number of successes in the n_0 historical data, $\mathbf{z}_{n_0}$. This distribution has mode s_0/n_0, the proportion of success in the historical data set. From standard conjugate analysis (Bernardo and Smith (1994), page 436) it follows that the marginal distribution of the sufficient statistics S is binomial–beta with parameters $(a_0 s_0 + 1, a_0(n_0 - s_0) + 1, n)$. Hence, closed form expressions are available, for instance, for the ALC.

5.3.1. Example

Greenhouse and Wasserman (1995) considered the problem of estimating the incidence θ of grade 4 leukopaenia, a life-threatening disease which is associated with an innovative therapy for mucositis (see Korn *et al.* (1993) for details). Suppose that we are interested in planning a new experiment for estimation of θ. Greenhouse and Wasserman (1995) reported results from a previous randomized trial, in which 12 out of 176 patients who received a similar treatment developed grade 4 leukopaenia. In using this information for setting up a design prior, we must take into account that the earlier study was historical and included previously untreated patients with different cancers from those of the new study. Hence, data from the earlier and the new study cannot be considered exchangeable. This motivates the need to discount evidence from the previous study. Let us set $n_0 = 176$ and $s_0 = 12$ and consider several values of the penalizing coefficient a_0.

As an example, we search minimal sample sizes which guarantee the width of the 95% equal-tails interval to be less than $k = 0.2$ (the ALC). Fig. 5 shows optimal sample sizes that were

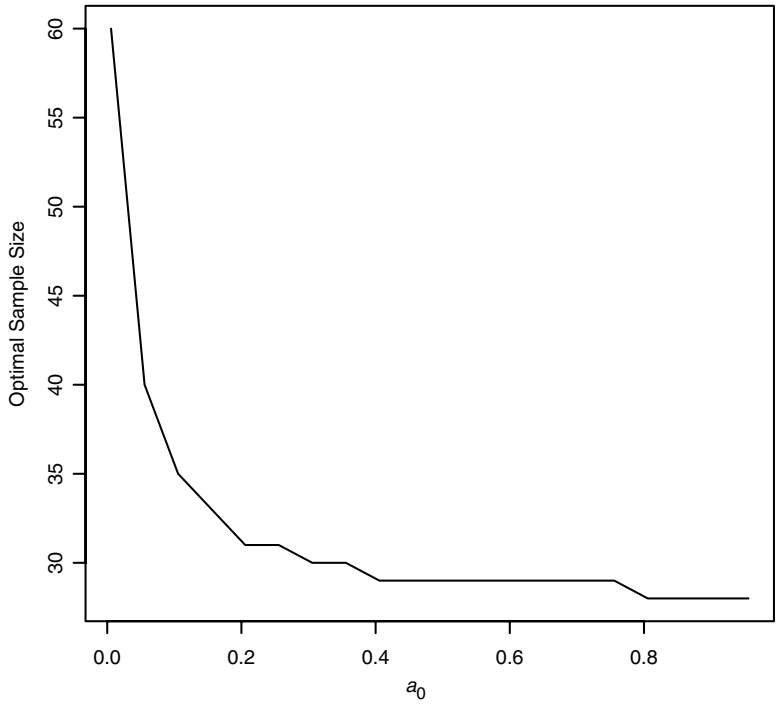

Fig. 5. Optimal sample size as a function of a_0, using ALC for 95% equal-tails intervals of the Bernoulli parameter ($n_0 = 176$ and $s_0 = 12$)

obtained for values of a_0 in the interval (0,1]. These values range from 60, when prior information is assigned the weight of one single observation ($a_0 = 1/n_0$), to 20, when full weight is given to the historical data ($a_0 = 1$). From Fig. 5 we can see that the reduction in sample size is quite substantial, passing from $1/n_0 = 0.006$ to values of a_0 which still denote severe discount of historical evidence. The decrease in optimal values of n becomes increasingly less dramatic as a_0 varies from 0.2, say, to 1. For instance, for $a_0 = 0.1$ and $a_0 = 0.2$ the sample sizes required are respectively equal to 35 and 31. In this example, for instance, a reasonable choice for a_0 seems to be $a_0 = 0.1$, since it allows the achievement of two goals: a serious discount of historical data and a substantial reduction in the sample size that is obtained with the choice $a_0 = 1/n_0$.

6. Multiple-power priors: sample size determination for comparing proportions

In the previous sections we have considered the most basic data structure, namely IID observations. Of course, more complex data require, in general, an extension of the power prior with a single discount parameter. As a motivating example, consider two alternative treatments for a given disease and suppose that we are interested in SSD for estimating the difference in success rates that are associated with the two therapies. Suppose also that historical information on the two treatments is unbalanced, knowledge on therapy 1 being substantially more accurate and reliable than on therapy 2. This circumstance requires us to apply distinct levels of discount to the two branches of historical information and makes the use of the standard power prior, with a unique discount parameter, intuitively inadequate. The point is that a severe discount of reliable historical data might yield unnecessarily large sample sizes and also an insufficient discount might yield too small a sample size. As shown in what follows, this problem can be addressed by using a more flexible multiple-fractions power prior. For the use of multiple-fractions power priors in posterior inference see, for instance, Brophy and Joseph (1995), Fryback *et al.* (2001) and Spiegelhalter *et al.* (2004).

The problem can be formalized as follows. Let $\mathbf{x}_{n_j} = (x_{j1}, \ldots, x_{jn_j})$, $j = 1, 2$, be independent data such that, given θ_j, x_{ji} are IID Bernoulli random variables. Suppose that we want to choose the number of observations n_1 and n_2 to make inference on the difference $\theta = \theta_1 - \theta_2$. Bayesian SSD for this problem has been previously studied, for instance, by Joseph *et al.* (1997) and Pham-Gia and Turkkan (2003), who also provided further references. Assuming the standard non-informative prior $\pi^{\mathrm{N}}(\theta_1, \theta_2) = 1$, it follows that, for sufficiently large sample sizes, the posterior distribution of θ is approximated by a normal density of parameters $(\mu_{s_1,s_2}, \sigma^2_{s_1,s_2})$, where

$$s_j = \sum_{i=1}^{n_j} x_{ji},$$

$$\mu_{s_1,s_2} = \sum_{j=1}^{2} \frac{s_j + 1}{n_j + 2},$$

$$\sigma^2_{s_1,s_2} = \sum_{j=1}^{2} \frac{(s_j + 1)(n_j - s_j + 1)}{(n_j + 2)^2 (n_j + 3)}.$$

As an SSD criterion, let us consider the ALC. Using the normal distribution approximation, the random length of the HPD set for θ, as a function of S_1 and S_2, is $L_\alpha(\mathbf{X}_{n_1}, \mathbf{X}_{n_2}) = 2z_{1-\alpha/2}\sigma_{s_1,s_2}$. To obtain a design prior for (θ_1, θ_2), suppose that we have independent historical data $\mathbf{z}_{n_0} =$

$(\mathbf{z}_{n_{0_1}}, \mathbf{z}_{n_{0_2}})$. Extending the basic definition, the power prior for this problem is proportional to

$$\pi_0(\theta_1, \theta_2) \prod_{j=1}^{2} L(\theta_j; \mathbf{z}_{n_{0j}})^{a_{0_j}}, \qquad a_{0_j} \in (0,1), \quad j = 1, 2,$$

where $L(\theta_j; \mathbf{z}_{n_j})$ is the binomial likelihood of θ_j that is associated with $\mathbf{z}_{n_{0j}}$. Using $\pi_0(\theta_1, \theta_2) = 1$, the power prior is proportional to the product of two independent beta distributions for θ_1 and θ_2, of parameters $(a_{0j}s_{0j} + 1, a_{0j}(n_{0j} - s_{0j}) + 1)$, $j = 1, 2$, where s_{0j} is the sum of the observations of $\mathbf{z}_{n_{0j}}$, $j = 1, 2$. Hence, the resulting distribution for θ_1 and θ_2 has the form of a power prior with multiple fractions, a_{0j}. As in the simple case, a_{0j} has the interpretation of the weight that we want to attach to the historical data from group j and $a_{0_1}n_{0_1} + a_{0_2}n_{0_2}$ plays the role of the 'prior sample size'. Analogously with the simple case, if $a_{0_1} = a_{0_2} = 1$, historical information is given full weight and different choices correspond to different degrees of discount. Given the above design prior, the marginal distribution of the sufficient statistics (S_1, S_2) turns out to be beta–binomial of parameters $(a_{0_j}s_{0_j} + 1, a_{0_j}(n_{0_j} - s_{0_j}) + 1, n_j)$, which allows analytical computations for the ALC.

6.1. Example

Joseph *et al.* (1997), pages 776–777, considered a clinical trial which had been planned to study the rates of myocardial infarction for patients who were affected by acute unstable *angina pectoris* and who followed two alternative study regimens. A previous study had shown that both aspirin and a combination of aspirin and heparin have the effect of lowering myocardial infarction rates which, in the study, turned out to be 4/121 and 2/122 respectively. Using this information, suppose that we want to determine the minimal sample size that is necessary to have a 0.95 HPD set for the effects difference θ, with length not greater than 0.05. As in Joseph *et al.* (1997), let us consider a balanced design, assuming $n_1 = n_2 = n$. Suppose also that the historical data for treatment 1 are judged to be homogeneous with data from the new experiment, whereas a considerable discount for historical data for therapy 2 is in order. This fact motivates the necessity of different treatments of the two groups of historical information. The goal of this example is to show that the use of the standard power prior with a single discount fraction may determine unnecessary severe discount of historical knowledge, with the consequence of yielding sample size values that are larger than needed. This problem is avoided by the more flexible multiple-power prior. Even though it would be straightforward to obtain plots of the ALC as a function of n for several choices of a_{0_1} and a_{0_2}, for brevity we here report only the expected length of the posterior interval for θ for a few values of the discount fractions and for a reference sample size, namely $n = 822$, that were reported by Joseph *et al.* (1997) as the frequentist optimal sample size. Table 2 shows the expected lengths of the 0.95 equal-tail sets for θ, at some chosen values of a_{0_j}, $j = 1, 2$.

The first row of Table 2 corresponds to the full use of historical data; the last row to severe discount. Intermediate rows correspond to unequal discounts. As expected, the stronger the overall weight that is given to historical data, the shorter is the corresponding expected size of the interval. More interestingly, if the same severe discount $a_{0_1} = a_{0_2} = 0.1$ is applied to both arms of historical data, i.e. if the standard power prior with $a_0 = 0.1$ is used, the expected length of the HPD set would be larger than the chosen threshold, 0.05, and 822 would then be considered insufficient. Conversely, using different discount factors, and choosing more severe values for a_{0_2} than for a_{0_1}, as appropriate in this problem, the average length of the HPD interval remains below 0.05 and 822 observations can be considered adequate.

Table 2. Intervals' average lengths for various a_{0_1} and a_{0_2}†

a_{0_1}	a_{0_2}	$E[L_\alpha(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})]$
1	1	0.0340
1	0.5	0.0357
0.5	0.5	0.0371
1	0.1	0.0444
0.1	0.1	0.0524

†$n_1 = n_2 = n = 822$.

7. Final remarks

The paper considers the use of historical data for Bayesian SSD. Specifically, we deal with the problem of how to use past evidence for the construction of priors in pre-experimental sample size calculations. We propose to use the power prior approach for two reasons:

(a) it is a partially automatic, easy-to-implement method;
(b) it allows discount of historical data, which is needed for taking into account a possible lack of homogeneity with data from the forthcoming experiment.

The resulting approach belongs to the category of hybrid classical–Bayesian methods for SSD, which make use of inferential tools that are acceptable both from the frequentist and the Bayesian viewpoint. See Spiegelhalter *et al.* (2004).

In the examples we have focused on the effect of levels discount of historical data on the optimal sample size. For all the examples that we have used, it is shown that, the lower the discount, the smaller is the resulting sample size that is required for the new experiment. However, it should be emphasized that the choice of a fraction that does not appropriately discount historical evidence might lead to an unrealistically small and thus inadequate sample size for the new experiment. Conversely, an excessive discount might determine the selection of an unnecessarily large sample size.

In general, the choice of the discount fraction a_0 is central for practical implementation of the power prior methodology. As pointed out in Section 3.1, there is no general unique answer to the problem. In the same section an example of classification of values for a_0 in terms of the severity of discount is given. However, from a practical point of view, plots of optimal sample size as a function of the discount factor can be quite informative. These plots permit visualization of the implications for sample size levels of different discounts, and eventually the choice of a sample size that is compatible with the degree of reliability that is attached to historical data.

The applied contexts that were used for exemplification in this paper are from the medical literature. However, there are many other applied fields, ranging from economics to reliability and quality control problems, in which we may expect to have pre-trial information that could be used for construction of the prior and that might need to be discounted.

Most of the examples that were considered in the paper deal with IID data, which are structurally simple. In these cases the single-fraction power prior has several justifications and might be considered sufficiently adequate for operating a discount of historical evidence. However, the approach needs refinements as soon as we move to a slightly more complex data structure, as shown in the still simple set-up of Section 6, where the exchangeability assumption is dropped. One task for future research is to extend the procedure to more complex SSD problems that

entail more complex models and likelihood structures. This is just an example that, however, suggests the necessity of an extension of the basic power prior definition. In this regard an aspect that deserves attention is the use of alternative methods for updating a starting prior with historical data. Specifically, it will be interesting to explore connections with the literature on likelihood and prior construction methods based on training samples. These topics are discussed, among others, by De Santis and Spezzaferri (2001), Ghosh and Samanta (2002) and Berger and Pericchi (2004).

Acknowledgements

The author is very grateful to the Joint Editor and to a referee for their helpful comments and suggestions. This research was partially supported by the University of Rome "La Sapienza" and by Ministero dell'Istruzione, dell'Università e della Ricerca Programmi di Ricerca 2003, Italy.

Appendix A: Explicit expressions for the average posterior variance, average length and length probability criteria for the example in Section 5.1

Under the assumptions of Section 5.1, it can be checked that

$$L(\mu, \lambda; \mathbf{z}_{n_0})^{a_0} \propto \lambda^{a_0 n_0/2} \exp(-\tfrac{1}{2} a_0 n_0 s_0^2 \lambda) \exp\{-\tfrac{1}{2} a_0 n_0 \lambda (\mu - \bar{z}_{n_0})^2\},$$

where $\bar{z}_{n_0}$ and n_0 are the sample mean and the size of the historical data and where s_0^2 is the (biased) sample variance. Hence, if $\pi_0 = \pi^{\mathrm{N}} \propto \lambda^{-1}$, it follows that $\pi^{\mathrm{P}}(\mu, \lambda | \mathbf{z}_{n_0}, a_0)$ is the product of a normal density for μ of parameters $(\bar{z}_{n_0}, a_0 n_0 \lambda)$ and a gamma density for λ of parameters $((a_0 n_0 - 1)/2, a_0 n_0 s_0^2/2)$. Let

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 / n$$

denote the (biased) sample variance. In this example both the posterior variance of μ and the length of credible intervals are functions of nS^2, whose marginal density is that of a gamma–gamma-distributed random variable of parameters $(\frac{1}{2}(a_0 n_0 - 1), \frac{1}{2} a_0 n_0 s_0^2, \frac{1}{2}(n-1))$. Computations of moments of gamma–gamma-distributed random variables, which correspond essentially to scale transformations of F-distributed variables, are straightforward. Hence, an explicit derivation of quantities for SSD is obtained by using standard calculations of conjugate analysis (see, for instance, Bernardo and Smith (1994), page 120, for details).

A.1. Average posterior variance criterion

It easy to check that the posterior variance of μ is $s^2/(n-3)$. This quantity depends on the data only through the sample variance and will be denoted as $\mathrm{var}(\mu | s^2)$. It can be shown that

$$\mathbb{E}[nS^2] = \frac{(n-1) a_0 n_0 s_0^2}{a_0 n_0 - 3}$$

and that

$$\mathbb{E}[\mathrm{var}(\mu | S^2)] = \frac{n-1}{n(n-3)} \frac{a_0 n_0 s_0^2}{a_0 n_0 - 3}, \qquad a_0 > \frac{3}{n_0}.$$

A.2. Average length criterion

Under the above prior assumptions, we have that

$$L_\alpha(\mathbf{X}_n) = 2 \frac{S}{\sqrt{(n-1)}} t_{n-1;1-\alpha/2},$$

where $t_{n-1;1-\alpha/2}$ is the $(1-\alpha/2)$-percentile of the t random variable with $n-1$ degrees of freedom. Then, it can be checked (see, for instance, Joseph and Belisle (1997), page 215) that

$$\mathbb{E}[L_\alpha(\mathbf{X}_n)] = 2t_{n-1;1-\alpha/2}\sqrt{\left\{\frac{a_0 n_0 s_0^2}{n(n-1)}\right\}}\frac{\Gamma(n/2)}{\Gamma\{(n-1)/2\}}\frac{\Gamma(\frac{1}{2}a_0 n_0 - 1)}{\Gamma\{\frac{1}{2}(a_0 n_0 - 1)\}}.$$

A.3. Length probability criterion

To compute the sample size according to the LPC, we need to compute, for a given desired length $l > 0$,

$$\Pr\{L_\alpha(\mathbf{X}_n) > l \mid a_0\} = 1 - F_{a_0}\left\{\frac{n(n-1)}{4t_{n-1;1-\alpha/2}^2}\, l^2\right\},$$

where $F_{a_0}(\cdot)$ is the cumulative distribution function of a gamma–gamma-distributed random variable of parameters $[\frac{1}{2}(a_0 n_0 - 1), \frac{1}{2}a_0 n_0 s_0^2, \frac{1}{2}(n-1)]$.

References

Adcock, C. J. (1997) Sample size determination: a review. *Statistician*, **46**, 261–283.

Berger, J. O. and Pericchi, L. R. (2004) Training samples in objective Bayesian model selection. *Ann. Statist.*, **32**, 841–869.

Bernardo, J. M. (1997) Statistical inference as a decision problem: the choice of sample size. *Statistician*, **46**, 151–153.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. New York: Wiley.

Brophy, J. M. and Joseph, L. (1995) Placing trials in context using Bayesian analysis: GUSTO revised by Reverend Bayes. *J. Am. Med. Ass.*, **273**, 871–875.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Statist. Sci.*, **10**, 237–308.

Clarke, B. S. and Yuan, A. (2006) Closed form expressions for Bayesian sample size. *Ann. Statist.*, **34**, in the press.

DasGupta, A. (1996) Review of optimal Bayes designs. In *Handbook of Statistics*, vol. 13, *Design and Analysis of Experiments* (eds S. Ghosh and C. R. Rao), pp. 1099–1147. New York: Elsevier.

De Santis, F. (2004) Statistical evidence and sample size determination for Bayesian hypothesis testing. *J. Statist. Planng Inf.*, **124**, 121–144.

De Santis, F. (2006a) Sample size determination for robust Bayesian analysis. *J. Am. Statist. Ass.*, **101**, 278–291.

De Santis, F. (2006b) Power priors and their use in clinical trials. *Am. Statistn*, **60**, 122–129.

De Santis, F. and Perone Pacifico, M. (2003) Two experimental settings in clinical trials: predictive criteria for choosing the sample size in interval estimation. In *Applied Bayesian Statistical Studies in Biology and Medicine* (eds M. Di Bacco, G. D'Amore and F. Scalfari). Norwell: Kluwer.

De Santis, F. and Spezzaferri, F. (1997) Alternative Bayes factors for model selection. *Can. J. Statist.*, **25**, 503–515.

De Santis, F. and Spezzaferri, F. (2001) Consistent fractional Bayes factor for nested normal linear models. *J. Statist. Planng Inf.*, **97**, 305–321.

Desu, M. M., and Raghavarao, D. (1990) *Sample Size Methodology*. San Diego: Academic Press.

Fayers, P. M., Cuschieri, A., Fielding, J., Craven, J., Uscinska, B. and Freedman, L. S. (2000) Sample size calculations for clinical trials: the impact of clinician beliefs. *Br. J. Cancer*, **82**, 213–219.

Freireich, E. J., Gehan, E., Frei III, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., Moon, J. H., Gendel, B. R., Spurr, C. L., Storrs, R., Haurani, F., Hoogstraten, B. and Lee, S. (1963) The effect of 6-Mercatopurine on the duration of steroid-induced remissions in acute leukemia: a model for evaluation of other potentially useful therapy. *Blood*, **21**, 699–716.

Fryback, G. D., Chinnis, Jr, J. O. and Ulvila, J. W. (2001) Bayesian cost-effectiveness analysis. *J. Technol. Assessmnt Hlth Care*, **17**, 83–97.

Ghosh, J. K. and Samanta, T. (2002) Nonsubjective Bayes testing—an overview. *J. Statist. Planng Inf.*, **103**, 205–223.

Greenhouse, J. B. and Wasserman, L. (1995) Robust Bayesian methods for monitoring clinical trials. *Statist. Med.*, **14**, 1379–1391.

Ibrahim, J. G. and Chen, M. H. (2000) Power prior distributions for regression models. *Statist. Sci.*, **15**, 46–60.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001) *Bayesian Survival Analysis*. Berlin: Springer.

Joseph, L. and Belisle, P. (1997) Bayesian sample size determination for normal means and differences between normal means. *Statistician*, **46**, 209–226.

Joseph, L., du Berger, R. and Belisle, P. (1997) Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist. Med.*, **16**, 769–781.

Joseph, L., Wolfson, D. B. and du Berger, R. (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *Statistician*, **44**, 143–154.

Julious, S. A. (2004) Tutorial in biostatistics: sample sizes for clinical trials with normal data. *Statist. Med.*, **23**, 1921–1986.
Korn, E. L., Yu, K. F. and Miller, L. L. (1993) Stopping a clinical trial very early because of toxicity: summarizing the evidence. *Contr. Clin. Trials*, **14**, 286–295.
Lindley, D. V. (1997) The choice of sample size. *Statistician*, **46**, 129–138.
O'Hagan, A. (1995) Fractional Bayes factors for model comparison (with discussion). *J. R. Statist. Soc.* B, **57**, 99–138.
O'Hagan, A. and Forster, J. (2004) *Bayesian Inference*, 2nd edn. London: Arnold.
Pham-Gia, T. and Turkkan, N. (1992) Sample size determination in Bayesian analysis. *Statistician*, **41**, 389–397.
Pham-Gia, T. and Turkkan, N. (2003) Determination of exact sample sizes in the Bayesian estimation of the difference of two proportions. *Statistician*, **52**, 131–150.
Raiffa, H. and Schlaifer, R. (1961) *Applied Statistical Decision Theory*. Boston: Harvard University Graduate School of Business Administration.
Sahu, S. K.and Smith, T. M. F. (2004) On a Bayesian sample size determination problem with applications to auditing. *Technical Report*. School of Mathematics, University of Southampton, Southampton.
Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004) *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. Chichester: Wiley.
Spiegelhalter, D. J. and Freedman, L. S. (1986) A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statist. Med.*, **5**, 1–13.
Walker, S. G. (2003) How many samples?: a Bayesian nonparametric approach. *Statistician*, **52**, 475–482.
Wang, F. and Gelfand, A. E. (2002) A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statist. Sci.*, **17**, 193–208.
Weiss, R. (1997) Bayesian sample size calculations for hypothesis testing. *Statistician*, **46**, 185–191.