Avoiding the Range of Equivalence in Clinical Trials: Bayesian Sample Size Determination for Robust Credible Intervals

Pierpaolo Brutti and Fulvio De Santis [1]

Università di Roma "La Sapienza"

Abstract

This article considers sample size determination methods based on Bayesian credible intervals for θ, an unknown real-valued parameter of interest. We assume that credible intervals are used to establish whether θ belongs to an indifference region. This problem is typical in clinical trials, where θ represents the effect-difference of two alternative treatments and experiments are judged conclusive only if one is able to exclude that θ belongs to a range of equivalence. Following a robust Bayesian approach, we model uncertainty on prior specification by a class Γ of distributions for θ and we assume that the data yield *robust evidence* if, as the prior varies in Γ, either the lower bound of the inferior limit of the credible set is sufficiently large or the upper bound of the superior limit is sufficiently small. Sample size determination criteria proposed in the article consist in selecting the minimal number of observations such that the experiment is likely to yield robust evidence. These criteria require computations of summaries of the predictive distributions of upper and lower bounds of the random limits of credible intervals. The method is developed assuming a normal mean as the parameter of interest and using conjugate priors. An application to the determination of sample size for a trial of surgery for gastric cancer is also illustrated.

Keywords: Bayesian power; Bayesian robustness; clinical trials; evidence; predictive analysis; sample size determination; superiority trials.

1 Introduction

Interval estimates of θ, an unknown scalar parameter of a statistical model, are commonly used for two different purposes: either for estimation or for testing hypotheses. When intervals are used for estimation, one typically wants them to be as short as possible. Conversely, if intervals are employed for testing, their length is not as relevant as their location. For instance, we are often interested in establishing the sign of θ, as it is the case when θ denotes the difference between two unknown quantities under comparison, and we want to establish whether one is larger than the other. In this circumstance, the most relevant information

[1] **Address for correspondence**: Dipartimento di Statistica, Probabilità e Statistiche Applicate. Università di Roma "La Sapienza". Piazzale A. Moro, 5 - 00185, Roma - Italy. **e.mail**: pierpaolo.brutti@uniroma1.it - fulvio.desantis@uniroma1.it

in an interval estimate is whether it contains or not zero or, in general, if it has intersection with an "indifference interval".

Even though the above mentioned problem and the methods proposed in this article applies to a variety of different contexts, clinical trials offer an ideal example of application for what we are going to deal with. Hence, in what follows we will often refer to this specific context and will specialize terminology to this setup. In this scenario, θ may represents improvement from a new treatment with respect to a standard therapy. A standard example is given by superiority trials, (see, for instance, Julious, 2004), in which the effects of two alternative treatments are compared in order to establish whether one is better than the other. In this case, an experiment is informative if the interval estimate for θ has no intersections with an indifference interval that, in this context is referred to as *range of equivalence*. Of course, depending on the values chosen for the range of equivalence and on the formulation of the problem, the above setup applies also to equivalence and non-inferiority trials (Julious, 2004).

The topic of this article is sample size determination (SSD) for interval estimation of θ. Depending on the inferential approach adopted and on the purpose intervals are used for, different SSD criteria can be used. Many criteria currently available are designed for the estimation use of intervals and aim to the pre-experimental control of either their length (for fixed coverage) or of their coverage (for fixed length): see, for instance, Armitage, Berry and Matthews, (2002), and, in the Bayesian framework, Joseph and Belisle (1997), Joseph, du Berger and Belisle (1997), De Santis and Perone Pacifico (2003). See also Adcock (1997) and Wang and Gelfand (2002) for reviews. In this article we specifically consider Bayesian SSD methods, under the assumption that interval estimates are used for testing goals. The basic idea is to choose the minimal number of observations so that range of equivalence and observed credible interval do not overlap, i.e. so that the inferior (superior) limit of the interval estimate is larger (smaller) than the superior (inferior) limit of the indifference region.

This problem has been already addressed, in the context of Bayesian design and analysis of clinical trials, by Spiegelhalter and Freedman (1986), who proposed a predictive approach to the sample size problem, based on careful elicitation and use of subjective clinical opinion. Their approach was to determine the sample size using the predictive probability of reaching a firm conclusion in favor of a treatment, i.e. of observing a lower (upper) limit of the interval estimate larger (smaller) than a value of "minimal clinically important difference". Under the usual setup, in which a normal mean represents the unknown improvement of a new treatment, they used classical confidence intervals (i.e. credible intervals determined with the standard noninformative constant prior) and proposed to evaluate the probability of firm

conclusions using, instead of the standard sampling distribution, the predictive distribution of the data. This was determined with a prior elicited using subjective clinical opinion. Their approach resulted in the definition of an average power function, to be used for sample size choice as an alternative to the standard frequentist power function. Extensions to the use of credible sets determined with proper priors are considered in Spiegelhalter et al. (2004).

In this paper we move from Spiegelhalter and Freedman's approach but we adopt instead a robust Bayesian viewpoint: we assume that, in place of a single distribution, a class of priors for the unknown parameter is considered. For reviews and references on the robust Bayesian viewpoint, see Berger (1984, 1990), Berger, Rios Insua and Ruggeri (2000) and Wasserman (1992). Given an observed sample, for each prior in the class one can virtually compute the limits of a $(1-\alpha)$-level credible interval and determine the lower and the upper bounds respectively of the inferior and the superior limits, as the prior varies in the class. Then, one can claim that the experiment yields *robust evidence* in favor of the hypothesis that the parameter is outside the range of equivalence (or, with Spiegelhalter and Freedman's terminology, that one has reached firm conclusions) if one of the following outcomes is observed: either the lower bound of the inferior limit of the interval estimate is sufficiently large or the upper bound of the superior limit is sufficiently small. In the design stage, the data, the posterior distribution and any of its functionals are random objects. Hence, credible sets as well as upper and lower bounds of their limits are also random. The idea is then simply to determine a sample size such that one has good chances of observing robust evidence. This entails computations with a marginal distribution of the data and formulation of criteria based on summaries of the predictive distributions of lower/upper bounds of the limits of the credible interval.

Motivations for the use of the robust Bayesian approach are given, for instance, in Berger et al. (2000). Essentially, the idea is that elicitation of a single prior is often affected by a considerable degree of uncertainty, that might be taken into account by replacing this single prior with a class of distributions. In design problems, it is of interest to determine sample sizes that are robust, to a certain extent, to uncertainty in the prior, i.e. that are able to guarantee the design goal for all the priors in a selected class of distributions. The importance of sensitivity/robustness studies has been often pointed out in the specific context of analysis of clinical trials data. Even though sensitivity checks should in principle concern both the likelihood and the prior, the Bayesian literature has focused mainly of the latter input. See, among others, Spiegelhalter et al. (2004, Section 5.6), Greenhouse and Wasserman (1995 and 1996), Carlin and Perez (2000), Carlin and Sargent (1996) and Sargent and Carlin(1996).

Robustness issues related to Bayesian interval estimation have been previously considered, for instance, by Berger and Berliner (1986) Pericchi and Walley (1991) and Wasserman

(1989). However, these papers are concerned on studying and comparing robustness of the posterior probability of credible intervals determined with a specific prior, as the prior varies in specific classes. Conversely, we are here concerned in studying bounds of the limits of credible intervals, for a given class of priors.

All of the forgoing is also related to the wider area of robust Bayesian experimental design: see Chaloner and Verdinelli (1995) and DasGupta (1996) for general reviews. Problems of sample size determination for robust Bayesian analysis, very close in the spirit to the ones presented in this article, are considered in DasGupta and Mukhopadhyay (1994), Ianus (2000) and De Santis (2005).

This article is structured as follows. Section 2, formalizes a methodology for SSD when the goal of the experiment is to yield robust evidence in favor of a treatment. Section 3 develops the method under the assumption that θ is a normal mean and that conjugate priors (Section 3.1) are used. Comparisons with non-robust and noninformative Bayesian approaches to SSD are discussed in Section 4. Section 5 deals with the unknown variance case. An extensive example in the context of designing a clinical trial is given in Section 6. Finally, Section 7 contains a discussion.

2 Methodology

2.1 Preliminaries and choice of priors

Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be a random sample with each X_i having density $f(x|\theta)$ that depends on a real parameter θ, the unknown difference between two alternative treatments. We are interested in establishing whether θ belongs to the interval $\mathcal{I} = [\theta_I, \theta_S]$, the treatments range of equivalence. Adopting a Bayesian perspective, let π_A be the prior distribution of θ, $\mathbf{x}_n$ the observed data and

$$\pi(\theta|\mathbf{x}_n; \pi_A) = \frac{f_n(\mathbf{x}_n|\theta)\pi_A(\theta)}{m_A(\mathbf{x}_n; \pi_A)}$$

the corresponding posterior density, where $m_A(\mathbf{x}_n; \pi_A) = \int_\Theta f_n(\mathbf{x}_n|\theta)\pi_A(\theta)d\theta$ is the marginal or prior predictive distribution of the data and $f_n(\mathbf{x}_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$. For a reason that will shortly become clear, we will refer to π_A as the *analysis prior*. Let us assume for simplicity that the posterior density is unimodal and let $C_\alpha(\mathbf{x}_n; \pi_A)$ be the $(1-\alpha)$-level posterior credible interval for θ:

$$C_\alpha(\mathbf{x}_n; \pi_A) = [\ell_n(\mathbf{x}_n; \pi), u_n(\mathbf{x}_n; \pi_A)].$$

Typical examples of C_α are the $(1-\alpha)$-level highest posterior density intervals and equal-tails intervals.

For the problem at hand, the experiment yielding $\mathbf{x}_n$ provides conclusive inferential *evidence* against the hypothesis that θ does not belong to $\mathcal{I}$ if

$$\ell_n(\mathbf{x}_n; \pi_A) > \theta_S \qquad \text{or} \qquad u_n(\mathbf{x}_n; \pi_A) < \theta_I.$$

We are now interested in the pre-experimental problem of determining a sample size such that the chances that the data yield evidence are sufficiently high. Before observing the data, $\ell_n(\mathbf{X}_n; \pi_A)$ and $u_n(\mathbf{X}_n; \pi_A)$, are random variables. Frequentist SSD methods typically consider an initial guess $\tilde{\theta}$ on the unknown parameter and use the joint density $f_n(\cdot|\tilde{\theta})$ of $\mathbf{X}_n$ for pre-posterior computations. However, this approach fails to account for uncertainty on θ and yields designs that are only locally optimal. See, for instance, Chaloner and Verdinelli (1995) and Spiegelhalter and Freedman (1986) for discussion. In the Bayesian approach to the SSD problem one can take into account uncertainty on the guessed value of the parameter by introducing a *design prior* π_D and by replacing $f_n(\cdot|\tilde{\theta})$ with the marginal distribution of the data $m_D(\mathbf{x}_n) = \int_\Theta f_n(\mathbf{x}_n|\theta)\pi_D(\theta)d\theta$ for pre-posterior computations.

The distinction between design and analysis priors is a central aspect of the methodology. This approach has been proposed in several previous articles: among these, see for instance Wang and Gelfand (2002), Sahu and Smith (2004), Joseph, du Berger and Belisle (1997), De Santis (2005). See also Clarke and Yuan (2005). The underlying idea is that analysis and design priors serves, in the SSD-inference process, two different purposes. The analysis prior, π_A, expresses prior knowledge/uncertainty on θ that we want to take into account in posterior analysis. The design prior, π_D, describes a scenario and it serves to account for uncertainty on possible guessed values for θ in the design stage. The marginal distribution, m_D, represents the data-generator mechanism that incorporates automatically the uncertainty on the guessed value for θ formalized by π_D. Following Wang and Gelfand (2002), we say that π_D arises in a "what-if" spirit: *what* sample size is appropriate for reaching conclusive inferential results *if* we assume that θ lies in a specific subspace of the parameter space and is distributed according to π_D? The necessity of a distinction between analysis and design prior is specifically evident in clinical trials. One example is when, in planning an experiment for inference on θ, the design prior π_D is centered on a value greater than θ_S, the superior limit of the range of equivalence, that represents a plausible significative effect level we wish to assess. At the same time, one might desire to assume an analysis prior that expresses neutrality (centered on zero, say) and that is relatively noninformative, so to let the data to drive the analysis, as often required by regulatory agencies. In other words, one might want to design the experiment under optimistic expectations, but also to be as neutral as possible in reporting posterior results.

Summarizing, π_D formalizes design expectations, whereas π_A expresses prior opinions / attitudes towards the new treatment and the weight we want to attach to it in the posterior.

The distinction between these two priors supplies flexibility to the design process. It gives the researcher the chance of checking the effect of combinations of more or less optimistic choices of π_D (i.e. expectations) with more or less optimistic choices of π_A (i.e. prior inputs) on optimal sample size. Of course, the choice of π_A and π_D is crucial in the resulting sample size. Fayers et al. (2000) point out for instance that, if sample size calculations are driven by a trial design committee that includes a number of enthusiasts for the trial treatment, and if the opinion of a larger community of experts that includes skeptics is not adequately accounted for, the risk of serious underestimation of the sample size is high.

Before moving on, a technical point: whereas π_A can be a standard noninformative and improper prior, in order for the marginal m_D to exist, the design prior π_D, in general, has to be proper. As a matter of fact, the use of noninformative analysis priors is the approach followed in several previous works on Bayes SSD: see, for instance, Spiegelhalter and Freedman (1986), Joseph et al. (1997), Wang and Gelfand (2002), Spiegelhalter et al. (2004) and De Santis (2004).

In this article we follow an alternative approach. We suppose that uncertainty on the analysis prior is modelled through a class of priors and consider a robust Bayesian approach to SSD for the interval estimation problem sketched in the previous section. This is an intermediate choice between choosing a noninformative prior and using a single proper prior in posterior analysis. Specifically, assume that, instead of a fixed prior π_A for θ, we have a class of prior distributions, Γ_A. For a given sample $\mathbf{x}_n$, the values of $\ell_n(\mathbf{x}_n; \pi_A)$ and $u_n(\mathbf{x}_n; \pi_A)$ change as π_A varies in Γ_A. Let

$$L_n(\mathbf{x}_n) = \inf_{\pi_A \in \Gamma_A} \ell_n(\mathbf{x}_n; \pi_A) \qquad \text{and} \qquad U_n(\mathbf{x}_n) = \sup_{\pi_A \in \Gamma_A} u_n(\mathbf{x}_n; \pi_A) \tag{1}$$

be the observed lower and upper bounds of ℓ_n and u_n, obtained as π_A varies in Γ_A. We say that the data yields *robust evidence* against the hypothesis that $\theta \in \mathcal{I}$ if

$$L_n(\mathbf{x}_n) = \inf_{\pi_A \in \Gamma_A} \ell_n(\mathbf{x}_n; \pi_A) > \theta_S \qquad \text{or} \qquad U_n(\mathbf{x}_n) = \sup_{\pi_A \in \Gamma_A} u_n(\mathbf{x}_n; \pi_A) < \theta_I,$$

i.e. if, for any prior in Γ_A, $C(\mathbf{x}_n; \pi_A) \cap \mathcal{I} = \emptyset$.

2.2 Robust Bayesian SSD criteria

Turning to the SSD problem, the choice of π_D determines two different alternative scenarios: one wants to design the experiment under the assumption that the guessed true value of θ is either larger than θ_S (scenario A) or smaller than θ_I (scenario B). We can formalize these two set-ups by choosing π_D centered on a value $\mu_D > \theta_S$ in the former case, and $\mu_D < \theta_I$, in the latter. In the following we will mainly consider scenario A, under which we are interested

in predictive control of $L_n(\mathbf{X}_n)$. Specifically, the requirement is now to be likely to observe data such that L_n is larger than θ_S, the upper limit of $\mathcal{I}$. Three summaries of the predictive distribution of this random variable and the corresponding sample size criteria are now listed.

1. Expectation criterion. The optimal sample size is

$$n_E^*(\Gamma_A) = \min\{n \in I\!N : \ e_n^L > \theta_S\},$$

where

$$e_n^L = \mathbb{E}_{m_D}[L_n(\mathbf{X}_n)] = \mathbb{E}_{m_D}[\inf_{\pi_A \in \Gamma_A} \ell_n(\mathbf{X}_n; \pi_a)]$$

and where $\mathbb{E}_{m_D}$ denotes the expected value with respect to the predictive density m_D.

2. Tail probability criterion. For a given $\epsilon \in (0, 1)$, the optimal sample size is

$$n_P^*(\Gamma_A) = \min\{n \in I\!N : \ p_n^L > \epsilon\},$$

where

$$p_n^L = \mathbb{P}_{m_D}[L_n(\mathbf{X}_n) > \theta_S]$$

and where $\mathbb{P}_{m_D}$ is the probability measure corresponding to the predictive density m_D.

3. Worst outcome criterion. The optimal sample size is

$$n_W^*(\Gamma_A) = \min\{n \in I\!N : \ w_n^L > \theta_S\},$$

where

$$w_n^L = \inf_{\mathbf{x}_n \in \mathcal{D}} L_n(\mathbf{x}_n)$$

and where $\mathcal{D}$ is a subset of the sample space for each element of which we want to be guaranteed that L_n is larger than θ_S. In the following as set $\mathcal{D}$ we will consider the subset $\mathcal{D}_\gamma$ of the sample space whose marginal density is greater than a value such that its predictive probability is equal to $1-\gamma$ (highest marginal density set, or $(1-\gamma)$-HMD set, in the following).

SSD criteria for scenario B are easily obtained, *mutatis mutandis*. Denoting with

$$e_n^U = \mathbb{E}_{m_D}[U_n(\mathbf{X}_n)], \qquad p_n^U = \mathbb{P}_{m_D}[U_n(\mathbf{X}_n) > \theta_I] \qquad \text{and} \qquad w_n^U = \sup_{\mathbf{x}_n \in \mathcal{D}} U_n(\mathbf{x}_n)$$

the three summaries of the predictive distribution of U_n, the corresponding optimal sample sizes are defined as

$$n_E^* = \min\{n \in I\!N : \ e_n^U < \theta_I\}, \qquad n_P^* = \min\{n \in I\!N : \ p_n^U < \epsilon'\}, \qquad \epsilon' \in (0, 1)$$

and

$$n_W^* = \min\{n \in I\!N : \ w_n^U < \theta_I\}.$$

The use of predictive expectation and tail-probability based criteria as well as of several alternative versions of the worst outcome criterion have been previously proposed in SSD problems aimed to the control of the length of credible intervals: see, for instance, Joseph and Belisle (1997) and references therein. Criteria 1-3 control three different aspects of the predictive distribution of L_n and typically lead to quite different optimal sample sizes. The expectation criterion is in principle the most straightforward (it does not require to choose ϵ, for instance) but, unlike criterion 2, it does not take into account variability of L_n. The worst outcome criterion requires a control on each data point in a relevant subset of the sample space and it leads to optimal sample sizes substantially larger than those found using the other methods, if the set $\mathcal{D}$ is large.

3 Sample size determination for robust inference of the normal mean

SSD criteria introduced in the previous section are now developed for interval estimation of the normal mean, using classes of conjugate priors. Whereas this is possibly the simplest example one can think of, it is still important for two reasons. First of all, normal models are widely used in clinical trials for their huge range of potential practical applications (see, for instance Spiegelhalter et al. (2004)). Second, the use of normal conjugate models yields explicit expressions for the three proposed criteria. This is quite a lucky situation, which allows us to study analytically the role of all the prior inputs on optimal sample sizes.

3.1 Conjugate analysis

Let us assume that X_i has normal density with unknown mean θ and known precision λ. Given two real and positive values n_A^L and n_A^U, such that $n_A^L < n_A^U$, consider the class of restricted conjugate priors, defined as follows:

$$\Gamma_{RC} = \left\{ \mathrm{N}(\theta|\mu_A, n_A\lambda); \quad n_A \in [n_A^L, n_A^U] \subset \mathbb{R} \right\}, \tag{2}$$

where n_A, the prior sample size, ranges between the values n_A^L and n_A^U and where, in general, $\mathrm{N}(\cdot|a, b)$ denotes the density function of a normal random variable of mean and precision (a, b). From standard conjugate analysis it follows that, for each prior in Γ_{RC}, the limits of the $(1-\alpha)$-level HPD density interval are

$$\ell_n(\mathbf{x}_n; n_A) = \frac{n\bar{x}_n + n_A\mu_A}{n + n_A} - z_{1-\frac{\alpha}{2}}[\lambda(n + n_A)]^{-1/2}$$

and

$$u_n(\mathbf{x}_n; n_A) = \frac{n\bar{x}_n + n_A\mu_A}{n + n_A} + z_{1-\frac{\alpha}{2}}[\lambda(n + n_A)]^{-1/2},$$

where z_ϵ denotes the ϵ-level percentile of the standard normal random variable. The following result provides the expression of the lower bound of ℓ_n and of the upper bound of u_n, as π_A varies in Γ_{RC}.

RESULT 1. *Assume that X_i has density $N(\cdot|\theta, \lambda)$, with λ known, $i = 1, \ldots, n$ and that π_A belongs to the class Γ_{RC} defined in (2). Then,*

$$L_n(\mathbf{x}_n) = \inf_{\pi_A \in \Gamma_{RC}} \ell_n(\mathbf{x}_n; n_A) = \begin{cases} \ell_n(\bar{x}_n; n_A^L) & \bar{x}_n < \mu_A + \xi_L \\ \ell_n(\bar{x}_n; n_A^*) & \mu_A + \xi_L < \bar{x}_n < \mu_A + \xi_U \\ \ell_n(\bar{x}_n; n_A^U) & \bar{x}_n > \mu_A + \xi_U \end{cases}$$

and

$$U_n(\mathbf{x}_n) = \sup_{\pi_A \in \Gamma_{RC}} u_n(\mathbf{x}_n; n_A) = \begin{cases} u_n(\bar{x}_n; n_A^U) & \bar{x}_n < \mu_A - \xi_U \\ u_n(\bar{x}_n; n_A^*) & \mu_A - \xi_U < \bar{x}_n < \mu_A - \xi_L \\ u_n(\bar{x}_n; n_A^L) & \bar{x}_n > \mu_A - \xi_L \end{cases},$$

where

$$\xi_k = \frac{z_{1-\alpha/2}}{2n}\left(\frac{n + n_A^k}{\lambda}\right)^{1/2}, \quad k = L, U, \qquad \textit{and} \qquad n_A^* = \frac{4n^2\lambda(\bar{x}_n - \mu_A)^2}{z_{1-\frac{\alpha}{2}}^2} - n.$$

PROOF. The explicit expression of L_n is obtained noting that, if $\bar{x}_n < \mu_A$, ℓ_n is an increasing function of n_A; whereas, if $\bar{x}_n > \mu_A$, ℓ_n has a minimum at n_A^*. The result follows by discussing the relative position of n_A^* with respect to the interval $[n_A^L, n_A^U]$. The expression of U_n is obtained similarly. □

Remarks

i) From the expression of L_n it can be checked that, as one intuitively expect, for values of μ_A sufficiently large (small), the lower bound of ℓ_n is attained in correspondence of the minimal (maximal) precision prior in Γ_{RC}, i.e. of the distribution with prior sample size equal to n_A^L (n_A^U).

ii) It is straightforward to check that, using the class Γ_C of conjugate priors with no restrictions on the variance, obtained from (2) as $n_A^L \to 0$ and $n_A^U \to +\infty$, the expression of $\inf_{\pi_A} \ell_n$ is

$$\begin{cases} \ell_n(\bar{x}_n; 0) & \bar{x}_n < \mu_A + \xi_L \\ \ell_n(\bar{x}_n; n_A^*) & \bar{x}_n > \mu_A + \xi_L \end{cases},$$

with $\xi_L = z_{1-\alpha/2}/(2\sqrt{n\lambda})$.

We are now interested in the predictive distribution of L_n and U_n but, for brevity, we here report results only for the former. Assume that the design prior for θ is

$$\pi_D(\theta) = N(\theta|\mu_D, \lambda n_D), \tag{3}$$

where μ_D can be interpreted as a guessed value for the unknown θ and where n_D denotes the prior sample size associated to π_D. It follows that the predictive distribution of $\bar{X}_n$ is

$$m_D(\bar{x}_n) = N(\bar{x}_n|\mu_D, \lambda_m), \qquad \lambda_m = \lambda(n^{-1} + n_D^{-1})^{-1}.$$

RESULT 2. *Under the assumptions of Result 1, using the design prior (3), the following results hold.*

a) The expression of $e_n^L = \mathbb{E}_{m_D}[L_n(\mathbf{X}_n)]$ *is*

$$e_n^L(\Gamma_{RC}) = \ell_n(\mu_D; n_A^L)\Phi(a_L) + \ell_n(\mu_D; n_A^U)[1 - \Phi(a_U)] + \mu_A[\Phi(a_U) - \Phi(a_L)] + H_n^L,$$

where

$$H_n^L = \frac{1}{\sqrt{2\pi}\lambda_m}[\psi_U e^{-\frac{1}{2}a_U^2} - \psi_L e^{-\frac{1}{2}a_L^2}] - \frac{z_{1-\alpha/2}^2}{4}(n\lambda)^{-1}\int_{\mu_A+\xi_L}^{\mu_A+\xi_U}(y-\mu_A)^{-1}N(y|\mu_D, \lambda_m)dy,$$

$$a_k = \sqrt{\lambda_m}(\mu_A - \mu_D + \xi_k), \qquad \text{and} \qquad \psi_k = \frac{n}{n + n_A^k}, \qquad k = L, U.$$

b) The expression of $p_n^L = \mathbb{P}_{m_D}[L_n > \theta_S]$, *for* $\mu_A < \theta_S$, *is*

$$\begin{aligned} p_n^L &= \left\{\Phi(a_L) - \Phi\left(\sqrt{\lambda_m}(b_L - \mu_D)\right)\right\} I_{(b_L,+\infty)}(\mu_A + \xi_L) \\ &+ \{\Phi(a_U) - \Phi(a_L)\} I_{(\theta_S,+\infty)}(\mu_A) \\ &+ 1 - \Phi\left[\sqrt{\lambda_m}(\max\{b_U, \mu_A + \xi_U\} - \mu_D)\right], \end{aligned}$$

where

$$b_k = \theta_S + n_A^k(\theta_S - \mu_A)/n + 2\xi_k, \qquad k = L, U.$$

c) Let $\mathcal{D}_\gamma \subset \mathbb{R}$ *be the* $(1-\gamma)$-HMD *interval. Then*

$$w_n^L = \inf_{\mathbf{x}_n \in \mathcal{D}_\gamma} L_n(\mathbf{x}_n) = L_n(\mu_D - z_{1-\frac{\gamma}{2}}\lambda_m^{-1/2}).$$

PROOF. Part a) follows from standard calculations, using the basic integral

$$\int_a^b y\mathrm{N}(y|m, v^{-2})dy = m\left[\Phi\left(\frac{b-m}{v}\right) - \Phi\left(\frac{a-m}{v}\right)\right] + \frac{v}{\sqrt{2\pi}}\left[e^{-\frac{1}{2}\left(\frac{a-m}{v}\right)^2} - e^{-\frac{1}{2}\left(\frac{b-m}{v}\right)^2}\right].$$

Part b) follows from standard probability calculations, using the expression of L_n derived in Result 1 and noting that, if $\mu_A < \theta_S$, the intersection of the events $(L_n > \theta_S)$ and $(\mu_A + \xi_L < \bar{X}_n < \mu_A + \xi_U)$ is empty. Part c) follows from noting that the $(1-\gamma)$-HMD interval is $\mathcal{D}_\gamma = (\bar{x}_n - z_{1-\frac{\gamma}{2}}\lambda_m^{-1/2}, \bar{x}_n + z_{1-\frac{\gamma}{2}}\lambda_m^{-1/2})$ and that L_n is a monotone increasing function of $\bar{x}_n$. Hence, as $\bar{x}_n$ varies in $\mathcal{D}_\gamma$, L_n attains the minimum at the lowest limit. □

Remarks

i) Letting $n_D \to +\infty$, one obtains the expressions corresponding to the use of the sampling distribution $N(\cdot|\mu_D, \lambda n)$ instead of the marginal distribution $N(\cdot|\mu_D, \lambda/(1/n + 1/n_D))$.

ii) The first and the second terms of e_n^L are respectively $O(1)$ and $O(n^{-1/2})$ whereas

$$[\psi_U e^{-\frac{1}{2}a_U^2} - \psi_L e^{-\frac{1}{2}a_L^2}] = O(n^{-1}), \quad (n\lambda)^{-1}\int_{\mu_A+\xi_L}^{\mu_A+\xi_U}(y-\mu_A)^{-1}N(y|\mu_D,\lambda_m)dy = O(n^{-3/2}).$$

Hence, for sufficiently large sample sizes, the leading terms in the expression of e_n^L are the first two.

iii) The quantity p_n^L represents the probability of reaching a robust significant evidence in favor of the hypothesis $\theta > \theta_S$ and can be interpreted as a robust Bayesian power. See Sections 3.1.2 and 4 for details.

The following Corollary describes the behavior of e_n^L, p_n^L and w_n^L as n goes to infinity.

COROLLARY. *Under the assumptions of Results 1 and 2, using the design prior (3) and assuming $\theta_S > \mu_A$, as n goes to ∞, the sequence of random variables $(L_n;\ n \in \mathbb{N})$ converges in law to a normal random variable of parameters $(\mu_D, \lambda n_D)$ and*

$$e_n^L \to \mu_D, \qquad p_n^L \to 1 - \Phi[\sqrt{\lambda n_D}(\theta_S - \mu_D)], \qquad w_n^L \to \mu_D - z_{1-\frac{\gamma}{2}}(\lambda n_D)^{-1/2}.$$

PROOF. The result follows noting that, as n goes to infinity, $\mathbb{P}_{m_D}[L_n < \theta_S]$ tends to $\Phi\left(\sqrt{\lambda n_D}(\theta_S - \mu_D)\right)$. □

The above limiting values of the three sequences e_n^L, p_n^L and w_n^L have to be taken into account when fixing the value θ_S so that the sample size problem is actually solvable. Note that, for any finite n_D, the sequence $\left(w_n^L;\ n \in \mathbb{N}\right)$ is definitively dominated by the sequence $\left(e_n^L;\ n \in \mathbb{N}\right)$. Note also, $\forall n_D \in \mathbb{N}$, the limit of the sequence $\left(p_n^L;\ n \in \mathbb{N}\right)$ is strictly less then one: it converges to one only if also $n_D \to +\infty$.

3.1.1 Expectation criterion

We now focus on Criterion 1 using the following numerical example.

EXAMPLE 1. Assume that $\lambda = 1$, $\mu_D = 3$, $n_D = 1$. Following the expectation criterion, we choose the minimal sample size such that e_n^L is larger than θ_S, the upper limit of the range of equivalence, that now we set equal to 2.5. Table 1 (columns 2-4) shows minimal sample sizes n_E^* computed for several values of n_A^L and n_A^U and for several values of the analysis prior mean, μ_A. As expected, one can notice what follows.

Table 1: Optimal sample sizes n_E^* for several classes Γ_{RC} and prior means μ_A ($\mu_D = 3$, $n_D = 1$, $\lambda = 1$, $\theta_S = 2.5$, $\alpha = 0.05$)

	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	n_0	n_0	n_0	π^N
μ_A	1 | 9	2 | 8	3 | 7	4 | 6	4	5	6	0
0	83 (83)	76 (76)	70 (70)	63 (63)	49	56	63	16
1.0	60 (60)	56 (56)	51 (51)	47 (47)	38	42	47	16
1.5	49 (48)	46 (45)	42 (42)	39 (39)	32	35	39	16
2.7	25 (18)	23 (18)	21 (17)	19 (17)	16	16	16	16
3.0	21 (15)	18 (14)	16 (13)	13 (12)	12	11	10	16

i) For any value of μ_A, the wider the class Γ_{RC} (i.e. the difference $n_A^U - n_A^L$) the larger the corresponding value $n_E^*(\Gamma_{RC})$. For instance, when $\mu_A = 0$, the restriction of the range of values for the prior sample size from [1,9] to [4,6] implies a reduction in minimal sample size of 20 units. On the other hand, for any specific choice of Γ_{RC}, the larger μ_A (i.e. the more optimistic the analysis prior), the smaller $n_E^*(\Gamma_{RC})$.

ii) The design prior sample size, n_D, which expresses the degree of uncertainty assigned to the guessed value μ_D, has a greater impact on robust optimal sample sizes for larger values of μ_A than for smaller values. The values in parentheses in Table 3.1.1 are determined for $n_D = \infty$, i.e. using the point mass design prior at μ_D (i.e. using $f_n(\cdot; \mu_D)$) in pre-posterior computations. Note that for $\mu_A = 0$ and $\mu_A = 1$, the optimal sample sizes are unchanged and that, as μ_A increases, the reduction in n_E^* is stronger and stronger. This can be explained by the fact that, for small μ_A, the lower bound of ℓ_n is achieved for relatively large values of n_A, making the impact of sensitivity to changes in variability of the sampling mean (i.e. of n_D), less important than they are for large values of μ_A. In this latter case, in fact, the lower bound of ℓ_n is attained for

small values of n_A, and changes in the variability of the sampling mean have a stronger impact on e_n^L and, eventually on n_E^*. This fact can also be appreciated in Figure 1, which shows e_n^L as a function of n, for several values of μ_A and of n_D, assuming $\lambda = 1$ and $n_A^L = 1$, $n_A^U = 9$. The first three curves from the top correspond to the optimistic value $\mu_A = 3$. The last two lines are for $\mu_A = 0$. The upper solid line is for $n_D = 1$, whereas the dashed-dotted and the dashed line correspond respectively to the cases $n_D = \infty$ and $n_D = 0.1$. In this case, for instance, for $n_D = \infty$, $n_E^* = 15$, that is 6 sample units less than the value found for $n_D = 1$. Consistently, as n_D gets smaller, n_E^* increases. Turning to the $\mu_A = 0$ case, the optimal n_E^* for $n_D = \infty$ does not differ from those at $n_D = 1$. A limited increase in optimal sample sizes is noticed for $n_D = 0.1$. $\Diamond$

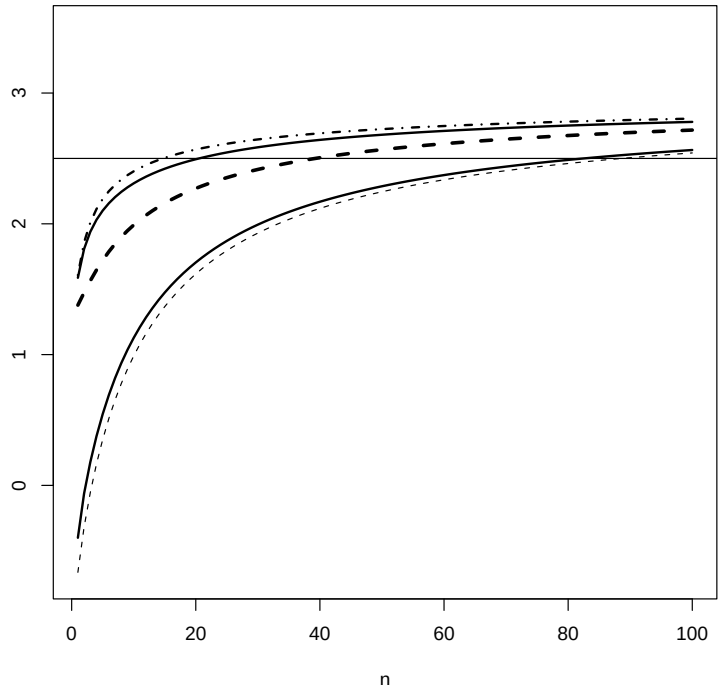

Figure 1: e_n^L for Example 1: $\mu_A = 0$, $n_D = 1$ (lower solid line); $\mu_A = 0$, $n_D = 0.1$ (dotted line); $\mu_A = 3$, $n_D = 1$ (upper solid line); $\mu_A = 3$, $n_D = \infty$ (dashed-dotted line); $\mu_A = 3$, $n_D = 0.1$ (dashed line);

The impact of the value chosen for the location of the design marginal, μ_D, is determinant. What we expect is that, the larger μ_D, the smaller the minimum n needed by e_n^L to reach the θ_S level. Recall that, in the clinical trial setting, μ_D expresses expectations that one has in planning the experiment, large values of it representing optimism towards the new treatment. As an example, Figure 2 reports the plots of e_n^L for $\mu_A = 0$, $n_A^L = 1$, $n_A^U = 9$, $n_D = 1$, $\lambda = 1$ and for three values of μ_D: 3 (solid), 2.7 (dashed) and 2.6 (dotted). The

horizontal line is drawn at the threshold level $\theta_S = 2.5$. From this plot one can appreciate the effect of μ_D on the steepness of e_n^L and, ultimately, on n_E^*'s, that result to be respectively equal to 83, 279 and 774. The effect of the precision, λ, is discussed in Section 5.

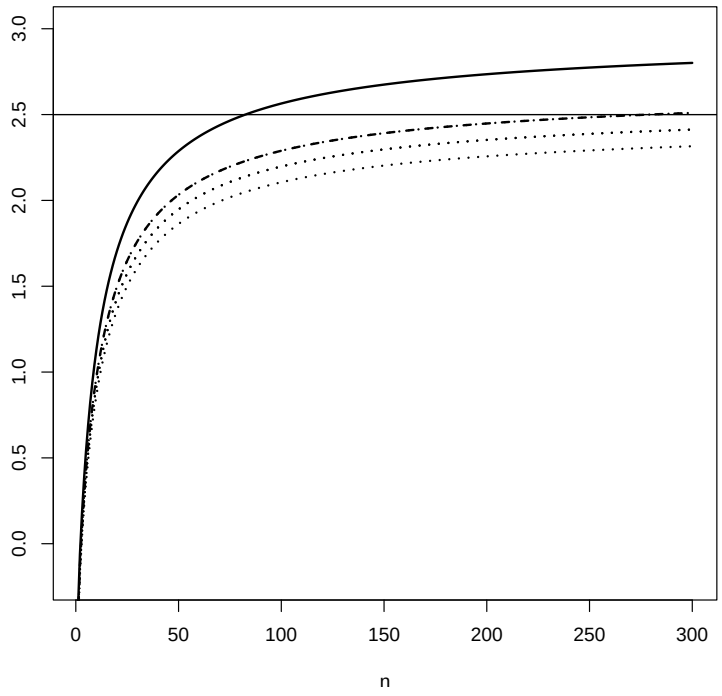

Figure 2: e_n^L for Example 1: $\mu_D = 3$ (solid line); $\mu_D = 2.7$ (dashed line) and $\mu_D = 2.6$ (dotted line);

3.1.2 Tail-probability criterion

As said above, the expectation criterion guarantees only an average control on the distribution of L_n, with no control on its variability. This might result in the selection of sample sizes for which the chances of reaching robust evidence are too small. For instance, for all the values n_E^* reported in Table 1 we have that $p^L_{n_E^*} \approx 0.5$. This is of course related to the asymptotic normality of L_n and to the specific values chosen for prior inputs, which imply the expected value and the median of L_n to get closer and closer. The use of the tail-probability criterion allows a stricter control on (the distribution) of L_n. Of course, this might require large sample sizes, depending on the input values and on the level ϵ. It is interesting to note that, unlike e_n^L, the asymptotic limit of p_n^L, i.e. the limiting value it can potentially reach, depends on λ, n_D and on the difference $\theta_S - \mu_D$. One can intuitively expect that, even for finite n, changes in these quantities have a substantial effect on p_n^L.

Example 2. Consider again the set-up of Example 1. Figure 3 reports the plots of p_n^L for for three distinct values of n_D: 1, 10 and $+\infty$, this third case corresponding to the use

of the point mass design prior on $\mu_D = 3$. For these three values, p_n^L is equal to 0.5 at $n_P^* = 83$. However, provided p_n^L is greater than 0.5, the larger is n_D, the greater is the value of p_n^L. For instance, the number of extra observations needed to increase p_n^L from 0.5 to 0.6 depends quite dramatically on n_D: only 6 extra observations when $n_D = +\infty$, 22 and 105 more when n_D is respectively equal to 10 and 1. This fact is relevant since it shows that ignoring uncertainty in the design - i.e. assuming a very large value for n_D, may lead to a non realistic assessment of the value attainable by p_n^L and of the sample size necessary for achieving conclusive evidence. For discussion and comparison between classical and Bayesian power, see also Spiegelhalter et al. (2004, Section 6.5). Table 2 (columns 2-5), reports the sample sizes n_P^* necessary to reach $\epsilon = 0.6$, for several values of n_A^L, n_A^U, μ_A and for the three values of n_D considered above. $\Diamond$

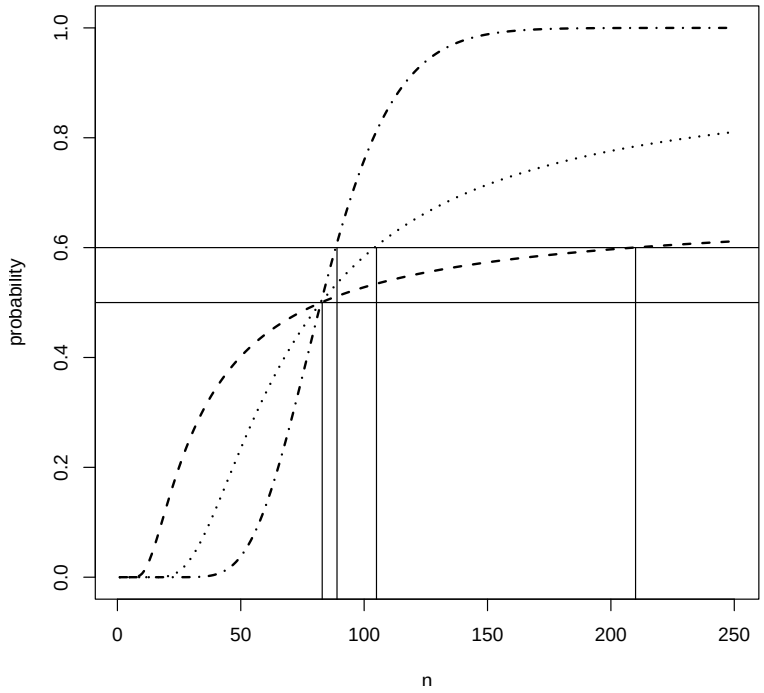

Figure 3: p_n^L for Example 2: $n_D = 1$ (dashed line); $n_D = 10$ (dotted line); $n_D = +\infty$ (dashed-dotted line).

3.1.3 Worst outcome criterion

It can be easily checked that w_n^L is an increasing function of γ that, as $\gamma \to 1$, tends to $L_n(\mu_D)$. For sufficiently large n, L_n is essentially linear in $\bar{X}_n$ and $L_n(\mu_D) = L_n(\mathbb{E}_{m_D}[\bar{X}_n]) \simeq \mathbb{E}_{m_D}[L_n(\bar{X}_n)] = e_n^L$. Therefore in this case we expect that, for any $\gamma < 1$, $w_n^L \leq e_n^L$ and hence that $n_W^* \geq n_E^*$ Of course, the smaller γ (i.e. the larger the set $\mathcal{D}_\gamma$), the larger the difference

Table 2: Optimal sample sizes n_P^* for $\epsilon = 0.6$ and for several classes Γ_{RC} and prior means μ_A ($\mu_D = 3$, $n_D = 1, 10, \infty$, $\lambda = 1$, $\theta_S = 2.5$, $\alpha = 0.05$)

	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	n_0	π^N
μ_A	1 | 9	2 | 8	3 | 7	4 | 6	5	0
0	210-105-89	195-97-82	181-89-75	166-80-68	151-72-61	65-24-20
1.0	158-77-65	149-71-61	140-66-56	130-61-52	120-55-47	"-"-"
1.5	132-62-53	125-58-50	118-55-46	111-51-43	104-47-40	"-"-"

between n_E^* and n_E^*. Note also that, as $n \to +\infty$, w_n^L tends to $w_\infty^L = \mu_D - z_{1-\frac{\gamma}{2}}(\lambda n_D)^{-1}$ whereas, as noticed above, e_n^L tends to μ_D. In the application of the proposed sample size criteria it is important to consider these limiting values in order to fix threshold values (here θ_S) that can actually be reached by e_n^L and w_n^L at sufficiently large but realistic sample size. This is numerically illustrated in the following example.

Example 3. Assume again that $\lambda = 1$, $\mu_D = 3$ and $n_D = 1$. We also assume that $\mu_A = 0$ and that $\theta_S = 1$. Table 3 reports the minimal sample sizes n_W^*, based on the worst outcome criterion, for values of $1 - \gamma$ ranging from 0.05 to 0.95. The last column of the table reports the asymptotic values w_∞^L that w_n^L can reach for the corresponding γ levels (recall that, in this case, $e_n^L \to 3$). The last row reports the values n_E^*, obtained using the expectation criterion. As expected, for any chosen γ level, n_W^* decreases as $n_A^U - n_A^L$ gets smaller. For any choice of $n_A^U - n_A^L$, the optimal sample size decreases as $1 - \gamma$ decreases. For very large values of $1 - \gamma$, implying a very strong predictive control on ℓ_n, the required sample size is very large. As an extreme example, for $1 - \gamma = 0.95$, the robust sample size is at least equal to 2740 observations. Note also that, as $1 - \gamma$ tends to zero, w_∞^L tends to the asymptotic value of e_n^L ($\mu_D = 3$) and the values n_E^* and n_W^* tend to get closer and closer. For completeness, we report in columns 7 and 8 of Table 3 the optimal sample sizes found using respectively a proper analysis prior ($n_0 = 5$) and the noninformative prior: similar comments to those given in Example 1 hold. $\Diamond$

4 Comparisons with non-robust and noninformative approaches

It is interesting to compare sample sizes obtained using the robust methods of the previous section to those determined with more traditional approaches based on the use of either a

Table 3: Optimal sample sizes n_W^* for several classes Γ_{RC} and several values of $1-\gamma$ ($\mu_A = 0$, $\mu_D = 3$, $n_D = 1$, $\lambda = 1$, $\theta_S = 1$, $\alpha = 0.05$)

		$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	n_0	π^N	w_∞^L
	$1-\gamma$	$1 \mid 9$	$2 \mid 8$	$3 \mid 7$	$4 \mid 6$	5	0	
n_W^*	0.95	2885	2835	2787	2740	2692	2446	1.040
	0.90	80	76	71	67	62	35	1.849
	0.75	25	23	22	20	18	7	2.325
	0.50	15	14	12	11	10	3	2.681
	0.05	9	9	8	7	6	2	2.940
	n_E^*	9	8	8	7	6	1	

specific proper prior or of a noninformative prior. We limit our analysis to the expectation criterion for ℓ_n under scenario A ($\mu_A \leq \mu_D$, $\mu_D > \theta_S$), in which we want to determine the smallest sample size such that e_n^L is sufficiently large. Similar considerations, omitted for brevity, can be extended to the worst outcome criterion.

Let us consider a specific prior π_0 for θ and let $e_n^L(\pi_0)$ denote the predictive expectation of ℓ_n determined using π_0. For any class of priors Γ such that $\pi_0 \in \Gamma$, and for any class Γ' such that $\Gamma \subset \Gamma'$ we have that $e_n^L(\Gamma') \leq e_n^L(\Gamma) \leq e_n^L(\pi_0)$ and hence that $n_E^*(\pi_0) \leq n_E^*(\Gamma) \leq n_E^*(\Gamma')$. For instance, in the specific case of Section 3, if $\pi_0 \in \Gamma_{RC} \subset \Gamma_C$, we have that

$$n_E^*(\pi_0) \leq n_E^*(\Gamma_{RC}) \leq n_E^*(\Gamma_C).$$

Table 1 shows the minimal sample sizes $n_E^*(\pi_0)$ obtained for some values of n_0 and allows us to quantify the number of extra observations implied by the robust approach with respect to the use of a single prior. Consider, for instance, $n_0 = 5$ (column 7). For any values of μ_A and for any class Γ_{RC}, $n_E^*(\pi_0) \leq n_E^*(\Gamma_{RC})$ but, however, as seen in Example 1, the impact of the value of μ_A is relevant: in fact, for $\mu_A = 0$ the relative reduction in optimal sample size that one has using π_0 (with $n_0 = 5$) in the place of Γ_{RC} (with $n_A \in [1, 9]$) is 32.5%, while, for $\mu_A = 3$, one has the 47.6% of reduction.

It is also interesting to compare sample sizes obtained with fixed analysis priors π_0, based on $e_n^L(\pi_0)$, for several values of n_0. Table 1 shows (columns 6-8) the values $n_E^*(\pi_0)$ for $n_0 = 4, 5, 6$. For small values of μ_A, as increasing weight is assigned to the prior centered on μ_A (i.e. as n_0 increases), the optimal sample size $n_E^*(\pi_0)$ increases as well; as the value of μ_A gets closer and closer to μ_D (here, for instance, for $\mu_D > 2.7$), $n_E^*(\pi_0)$ tends to decrease as n_0 increases. Again, the assignment of increasing weight to the prior mean μ_A results in a reduction of the minimal sample size required for making, on average, ℓ_n large only when

μ_A is sufficiently optimistic. Specifically, it can be easily checked that, for $\mu_A \geq \mu_D$, $e_n^L(\pi_0)$ increases monotonically with n_0 and that, as a consequence, $n_E^*(\pi_0)$ is a decreasing function of n_0.

Let us now consider the standard noninformative prior $\pi^N \propto 1$ that corresponds to assigning a null weight to the analysis prior mean, μ_A. It can be checked that

$$e_n^L(\pi^N) = \mu_D - z_{1-\frac{\alpha}{2}}(n\lambda)^{-1/2} \qquad \text{and} \qquad n_E^*(\pi^N) = \left\lceil \frac{z_{1-\frac{\alpha}{2}}^2}{\lambda(\mu_D - \theta_S)^2} \right\rceil,$$

where $n_E^*(\pi^N)$ is the smallest n such that $e_n^L(\pi^N) > \theta_S$. The optimal noninformative sample size $n_E^*(\pi^N)$ is a decreasing function of the sampling precision λ and of the difference $\mu_D - \theta_S$. Note that, as the difference $\mu_D - \theta_S$ becomes smaller and smaller, $n_E^*(\pi^N)$ diverges. In fact, in this case, $e_n^L(\pi^N)$ is strictly smaller than θ_S for any $n \in \mathbb{N}$. We now compare these sample sizes, based on π_N, to those determined using a proper analysis prior, which assigns weight n_0 to μ_A. Ideas carry over to the robust case. In general, it depends on the value of μ_A whether it is convenient to use π_0 or π^N, i.e. whether $n_E^*(\pi_0) < n_E^*(\pi^N)$ or $n_E^*(\pi_0) > n_E^*(\pi^N)$. Intuitively, under scenario A, we expect that for large values of μ_A (optimistic priors), it is convenient to assign as much weight as possible to μ_A, i.e. we expect $n^*(\pi_0) < n^*(\pi_N)$. As a limiting case, when $\mu_A = \mu_D$, it is straightforward to check that $e_n^L(\pi_0) = n\mu_D/(n + n_0 - z[\lambda(n + n_0)]^{-1/2})$ and that $n_E^*(\pi_0) = n_E^*(\pi^N) - n_0$. Hence n_0 is, in this case, the number of sample units saved using π_0 rather than π^N. Conversely, for small values of μ_A (skeptical prior), we expect to be convenient (in terms of minimal sample sizes) to use π^N rather than π_0, i.e. that $n_E^*(\pi_0) < n_E^*(\pi^N)$. Summarizing, we expect that, for sufficiently small values of μ_A, $n_E^*(\pi^N) \leq n_E^*(\pi_0) \leq n_E^*(\Gamma_{RC})$, whereas, for sufficiently large values of μ_A, $n_E^*(\pi_0) \leq n_E^*(\Gamma_{RC}) \leq n_E^*(\pi^N)$. Table 1 exemplifies numerically the above considerations: $n_E^*(\pi^N)$ is uniformly smaller than the corresponding sample sizes determined with both π_0 and Γ_{RC} for values of μ_A less than 2.7; when $\mu_A = \mu_D = 3$, $n_E^*(\pi^N)$ is always larger than $n^*(\pi_0)$ for the 3 values of n_0 considered and than $n_E^*(\Gamma_{RC})$ only for the two smallest classes considered.

Let us now turn to the tail-probability criterion. The quantity p_n^L is related to the classical concept of power and to the several versions of Bayesian power (Spiegelhalter et al. (2004, Section 5)), defined as probabilities of reaching "significant" results when testing an hypothesis on θ, in our notation, $\ell_n > \theta_S$. In fact, p_n^L can be interpreted as a form of robust Bayesian power and represents an extension of the *expected power* and of the *Bayesian power* reported in Speigelhalter et al. (2004). The expected power, defined as the predictive probability of observing a classically significant result, is essentially obtained from p_n^L by setting $n_A^L = n_A = U = 0$ (i.e. using a noninformative analysis prior). The Bayesian power, defined as the predictive probability that the posterior probability of a

certain hypothesis is larger than a given threshold, is related to the quantity obtained by p_n^L by setting $n_A^L = n_A^U = n_0$ (i.e. using a specific analysis prior instead of Γ_A). The differences between these concepts of power and the robust approach of the present paper are: *a*) the distinction between design and analysis priors; *b*) the use of a class of analysis priors in the place of either a noninformative prior or a single prior. The last two columns of Table 2 reports the values of n_P^* obtained respectively with the single analysis priors $N(\theta|\mu_A, n_0\lambda)$, where $n_0 = 5$, and the noninformative priors. From this table one can appreciate the increase in sample size due to the use of a class of analysis priors in the place of a single proper or noninformative priors i.e., the extent of sample size miscalculation due to ignoring uncertainty in the analysis prior.

5 Unknown variance

The analysis developed for the normal model relies on the restrictive assumption that λ is known. However optimal sample sizes might depend crucially on this parameter. A typical situation is depicted in Figure 4 where we see how, applying the expectation criterion of Section 2, the optimal sample sizes rapidly increase from 65 to 124 as λ ranges from 3 to $\frac{1}{3}$. This plot clearly shows how sensitive the proposed criteria might be to sampling precision and, consequently, suggests the necessity of using priors for λ that account for its uncertainty. In what follows we will consider only Criterion 1 and predictive expectation will now be denoted with $\widetilde{e}_n^L$. In the conjugate normal framework previously adopted, a natural choice is to consider an analysis prior $\pi_A(\theta, \lambda) = \pi_A(\theta|\lambda)\pi_A(\lambda)$, where $\pi_A(\theta|\lambda) \in \Gamma_{RC}$, defined in Equation 2 and where $\pi_A(\lambda)$ is a conjugate gamma density. From standard results, the posterior distribution of θ turns out to be

$$\pi(\theta|\mathbf{x}_n, n_A) = \mathrm{St}\left(\theta\middle|\theta_n, \tfrac{(n_A+n)(2\nu+n)}{2\beta+2g(\mathbf{x}_n)}, 2\nu+n\right),$$

where

$$\theta_n = \tfrac{n_A\mu_A+n\overline{x}_n}{n_A+n}, \quad g(\mathbf{x}_n) = \tfrac{1}{2}\left(nS^2 + \tfrac{n_A n(\overline{x}_n-\mu_A)^2}{n_A+n}\right),$$

S^2 is the sample variance, and $\mathrm{St}(a, b, c)$ denotes the density function of a Student t distribution with location, scale and degrees of freedom equal to (a, b, c). The extrema of the $(1-\alpha)$–level HPD density interval coincides with those of the equal tails interval and are given by

$$\widetilde{\ell}_n(\mathbf{x}_n; n_A) = \theta_n - t_{2\nu+n;1-\frac{\alpha}{2}}\sqrt{\tfrac{2\beta+2g(\mathbf{x}_n)}{(n_A+n)(2\nu+n)}}$$

and

$$\widetilde{u}_n(\mathbf{x}_n; n_A) = \theta_n + t_{2\nu+n;1-\frac{\alpha}{2}}\sqrt{\tfrac{2\beta+2g(\mathbf{x}_n)}{(n_A+n)(2\nu+n)}},$$

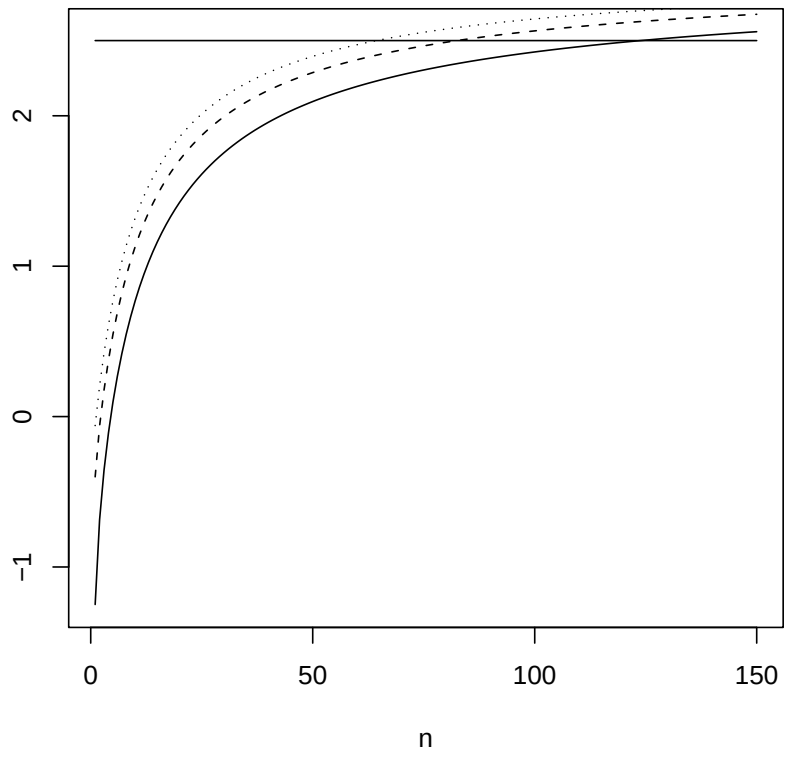

Figure 4: Sensitivity check with respect to variations of the prior precision λ ($\mu_A = 0, \mu_D = 3, n_D = 1, n_A^L = 1, n_A^U = 9$): e_n^L for $\lambda = \frac{1}{3}$ (solid line); $\lambda = 1$ (dashed line); $\lambda = 3$ (dotted line).

where $t_{\eta,\epsilon}$ denotes the ϵ–level percentile of a Student t distribution with η degrees of freedom. Although closed form expressions for $\widetilde{L}_n(\mathbf{x}_n) = \inf_{\pi_A \in \Gamma_A} \widetilde{\ell}_n(\mathbf{x}_n; \pi_A)$ and $\widetilde{U}_n(\mathbf{x}_n) = \sup_{\pi_A \in \Gamma_A} \widetilde{u}_n(\mathbf{x}_n; \pi_A)$ are still available, they are not structurally as neat as the previous ones, given the coupling between sample mean and variance that manifests itself in their explicit definitions. Nevertheless, numerical computations of the proposed criteria in this new setup are straightforward. Specifically, assuming $\pi_D(\theta, \lambda) = \pi_D(\theta|\lambda)\pi_D(\lambda)$, where $\pi_D(\theta|\lambda) = N(\theta|\mu_D, n_D\lambda)$ and where $\pi_D(\lambda) = \text{Ga}(\nu, \beta)$, we see that $m(x; \pi_D) = \text{St}(x|\mu_D, \omega_n^2, 2\nu)$, where $\omega_n^2 = [n_D(n_D + n)^{-1}\nu\beta^{-1}]^{-1}$.

Figures 5A and 5B show, as function of n and for the input values of Example 1, the plots of $\widetilde{e}_n^L$ computed with a gamma prior for λ, with parameters ν and β such that Mode$[\lambda] = 1$ and such that the prior probability of the set $(\frac{1}{3}, 3)$ is approximatively equal to 0.95. Comparing the new optimal sample sizes with those associated to the plots in Figures 1 and 2 where λ was fixed and equal to 1, we see how they seem to behave quite differently

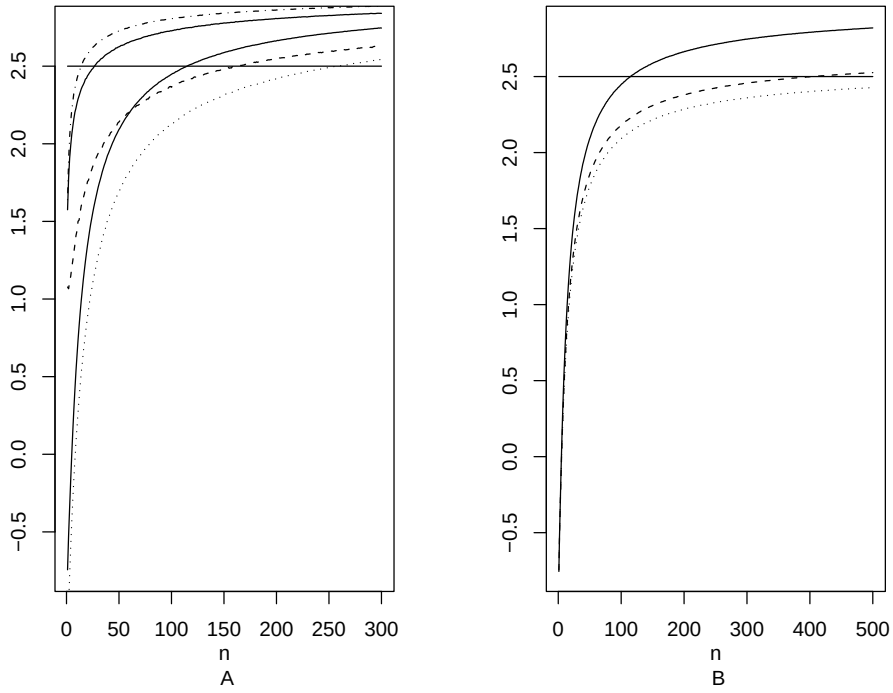

Figure 5: **A**: $\widetilde{e}_n^L$ for ($\mu_A = 0,\ n_D = 1$) (lower solid line); ($\mu_A = 0,\ n_D = 0.1$) (dotted line); ($\mu_A = 3,\ n_D = 1$) (upper solid line); ($\mu_A = 3,\ n_D = \infty$) (dashed-dotted line); ($\mu_A = 3,\ n_D = 0.1$) (dashed line). **B**: $\widetilde{e}_n^L$ for $\mu_D = 3$ (solid line); $\mu_D = 2.7$ (dashed line) and $\mu_D = 2.6$ (dotted line);

depending on the magnitude of other parameters like n_D: quite close to each other for large values of n_D ($n_E^\star = 15$ and $\widetilde{n}_E^\star = 14$, when $\mu_A = 3$ and $n_D = +\infty$); substantially far apart for small values of n_D ($n_E^\star = 40$ and $\widetilde{n}_E^\star = 164$, when $\mu_A = 3$ and $n_D = 0.1$).

6 An application to the design of a clinical trial

In this section we apply the proposed SSD methods in the context of planning the size of a clinical trial. In this regard, we revise the UK Medical Research Council (MRC) randomized trial of gastric surgery, ST01, discussed in Fayers et al. (2000) and Spiegelhalter et al. (2004, pp. 197-201). This trial was conducted from 1986 to 1994 in order to compare survival following conventional and radical surgery for gastric cancer. The goal was estimation of the log hazard ratio (log-HR), θ, values greater than zero favoring radical treatment. The classical statistic $Y_n = 4L_n/n$ was used, where L_n is the standard observed log-rank statistic and n the total number of events (deaths) observed. In planning the sample size, it has

been assumed that Y_n was approximately normally distributed, with parameters $(\theta, n\lambda)$. It was assumed that $\lambda = 1/4$ (for details and justifications, see Speigelhalter et al. (2004), pages 198). The sample size for the trial was determined according to the opinion of the surgical members of the design team which suggested that a change in 5-years survival from 20% (conventional) to 33.5% (radical) could be realistic and medically important. Hence, the number of units was determined so that the trial were able to detect a 13.5% difference (equivalent to a log-HR of 0.39) at the 5% significance level with 90% power. (Note: the value 0.39 will be used here as the design value, μ_D in our notation). This resulted in an overall sample size of 400 patients, predicted to yield $n = 276$ events. Based on the opinion of further 23 surgeons experienced in treating gastric carcinoma, Fayers et. al (2000) elicited a normal (analysis) prior distribution for θ, of mean 0.12, a value corresponding to 4% of improvement with respect to the baseline survival of the conventional therapy, and precision $n_0\lambda = (0.19)^{-1}$, so that $n_0 = 4(0.19^2) = 111$. (Note: these values will be used here to formalize the analysis prior class, Γ_A). Compared to the prior mean, the design value $\mu_D = 0.39$ appears as rather optimistic, and that it was found to yield a too small sample size. See Fayers et al. (2000) for discussion.

In addition to prior beliefs, Fayer et al. (2000) elicited also demands for radical surgery, and established that an improvement around 10% (which corresponds to a value of 0.29 on the log-HR scale) was judged necessary before switching to the new treatment, due to its impact and complication risks. (Note: this value can be reasonably used as superior limit of the range of equivalence, θ_S).

Let us now reconsider the sample size problem just illustrated and reformulate it in terms of the robust Bayesian approach based on predictive control of lower bounds of credible interval inferior limits. For brevity we here limit consideration to Criterion 1. Using the normal approximations for Y_n, results of the previous sections can be applied to this problem. As design prior, we consider a normal density with $\mu_D = 0.39$; several values for n_D will be used in numerical illustration. For the analysis prior mean, in addition to the clinical value $\mu_A = 0.12$ (4% improvement), we also consider the skeptical value $\mu_A = 0$ (no improvement) and the more optimistic value $\mu_A = 0.29$ (10% improvement). The class Γ_{RC} is defined by fixing $n_A^U = 111$, the full prior sample size elicited by Fayers et al. (2000), and $n_A^L = 1$. This choice corresponds to assigning to the weight of a single experimental unit to the prior, and yields the so-called unit-information prior (Kass and Wasserman, 1996). In order to explore the effect of the size of the class of priors (i.e. of the difference $n_A^U - n_A^L$), we also consider, just as an example, the values $n_A^L = 0.25 \times 111$ and $n_A^U = 0.75 \times 111$. Finally, we also include computation with the single prior obtained assuming $n_0 = 111/2$.

Implementation of the sample size method require to fix θ_S. Fayers et al. (2000) and

Spiegelhalter et al. (2004) point out that: *"... around a* 10% *of improvement was judged to be necessary before wishing to routinely implement the more radical surgery..."*. This corresponds to choosing $\theta_S = 0.29$. However, even under the rather optimistic design assumption of an expected improvement around 13.5 %, the requirement that the lower bound of the inferior limit of a credible interval be larger than this value results quite demanding and implies very large sample sizes. Hence, we here fix two more modest but more realistic values for the superior limit of the range of equivalence, namely, $\theta_S = 0.205$ and $\theta_S = 0.149$ corresponding respectively 7% and 5% improvement in survival of radical surgery versus conventional treatment. Table 6 reports the minimal sample sizes n_E^* based on the expectation criterion, for several values of μ_A, n_D, θ_S, for two classes Γ_A and for the selected single analysis prior, determined with design prior mean $\mu_D = 0.39$. As expected, sample sizes necessary to have the mean value of L_n greater than θ_S, decreases as μ_A and/or n_D increases and as $n_A^U - n_A^L$ decreases. In particular, the impact of the precision of the design prior n_D is quite relevant.

Table 4: Gastric Example. Optimal sample sizes n_E^*

		$\theta_S = 0.205$			$\theta_S = 0.149$		
		$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	n_0	$n_A^L \mid n_A^U$	$n_A^L \mid n_A^U$	n_0
μ_A	n_D	$1 \mid 111$	$\frac{1}{4}111 \mid \frac{3}{4}111$	$\frac{1}{2}111$	$1 \mid 111$	$\frac{1}{4}111 \mid \frac{3}{4}111$	$\frac{1}{2}111$
0	1	1301	968	607	875	631	370
	10	857	736	”	546	460	”
	100	748	680	”	463	418	”
0.12	1	1245	911	545	834	588	324
	10	785	668	”	494	411	”
	100	648	598	”	392	359	”
0.29	1	1170	830	452	778	527	254
	10	695	577	”	430	344	”
	100	537	496	”	318	287	”

In a second elicitation task, conducted when the trial was complete but before disclosure of the results and based again on the opinion of the trial committee, it was found that approximately 10% (log-HR=0.29) improvement was more realistic than 13.5% initially supposed. We have repeated calculations that yielded Table 6, replacing 0.39 with 0.29 as value for μ_D. The resulting sample sizes arising under this less optimistic scenario are, for all the cases considered in Table 6, uniformly much larger than the previous ones and, in general, unrealistic. For instance, for $\mu_A = 0.29$, $n_D = 100$, $\theta_S = 0.205$, for the two classes consid-

Table 5: Gastric Example. Optimal sample sizes $n_E^*(\pi^N)$

μ_D	$\theta_S = 0.205$	$\theta_S = 0.149$
0.39	449	265
0.29	2127	773

ered the optimal sample sizes are respectively equal to 2278 and 2177 (instead of 537 and 496 of the previous case). The conclusion is the following: if the goal of the experiment is so demanding that it requires an unrealistically large sample size for being achieved, either the trial is not even started, or demands are adequately reformulated. Whereas this is a general guideline in sample size determination problems, this is particularly relevant in this robust approach.

Finally, Table 6 reports the optimal sample sizes for the gastric example, when a noninformative analysis prior is used. As noted above, the resulting sample sizes do not depend neither on the analysis prior mean μ_A nor on n_D and are quite sensitive to the values chosen for both μ_D and θ_S. The noninformative approach, that neglects the role of μ_A and n_D, might lead to sample sizes that strongly differ from those obtained using either a proper analysis prior or the class Γ_A and, depending on the situations, might yield quite inadequate a sample size.

7 Discussion

In this article we propose SSD methods for interval estimation of a real-valued parameter, with specific focus on employment of the resulting criteria in clinical trials. The main characteristics of the proposed methodologies are: *a*) the use of a design prior for formalizing uncertainty on the guessed value of the parameter at the design stage of the inferential process; *b*) the distinction between analysis and design priors; *c*) the formalization of the uncertainty on elicitation of the analysis prior through the introduction of a class of distributions and, hence, of a robust approach. As far as *a*) is concerned, the use of a design prior qualifies the Bayesianity of the approach and has been previously acknowledged to be advantageous with respect to the standard practice of fixing a single value for the unknown parameter of the model. With respect to *b*), the formal proposal of using distinct priors for design and analysis is more recent in the literature (see Wang and Gelfand, 2002). This approach allows one to evaluate the impact on SSD of the mixing of different possible scenarios (represented by π_D) and prior expectations (represented by π_A). For instance one

can establish, in terms of required minimal sample sizes, the effect of combining optimism or skepticism - formalized by suitable choices for the parameters of the analysis prior - with more or less optimistic expectations on a new treatment, represented by the design prior. Finally, turning to *c*), the robust approach, which represent the major novelty of our contribution, allows designers of experiments to account for uncertainty in the elicitation of the prior distribution that describes pre-experimental information on the parameter. All the above results in a greater flexibility than the standard classical and Bayesian methods. As shown in all the numerical examples throughout the article, the price of such a greater flexibility is that the resulting sample sizes are in general larger than those obtained with classical or non robust Bayesian procedures. At a first look, this may be considered a drawback of the proposed robust Bayesian approach to the SSD problem. On the contrary, this approach avoids the risk that, ignoring uncertainty on the guessed value as well as on the elicited analysis prior, one ends up with unrealistically small sample sizes, which do not protect against an unsuccessful outcome.

In the article, in the context of normal models with classes of conjugate priors, we have compared different aspects of the predictive distributions of the lower bound of the inferior limit of credible intervals. Criterion 1, based on e_n^L is the easiest to employ, but it allows only a loose control on the distribution of L_n. Criterion 3, based on w_n^L, may determine a much stronger control on L_n, but it typically yields too large sample sizes. Criterion 2, based on p_n^L, is the most informative quantity among the three methods. The plots of all the three quantities e_n^L, p_n^L and w_n^L, as functions of n, are informative and allow to visualize the progressive gain one can achieve by increasing the sample size. This is particularly useful when one has to establish whether the gain in the chances of achieving robust evidence due to the increase in the size of the experiment is worth the cost of these units.

The approach presents some undeniable limits. Among these, one possible objection is that the class of conjugate priors is not large enough to represent properly and realistically uncertainty on the elicited analysis prior. This is certainly true. However this class presents also some non negligible advantages over more refined and flexible classes of prior distributions. The first advantage is analytical tractability, a characteristic that is not shared by other classes of priors, such as ϵ-contaminated distributions. Furthermore, one can hope that conjugate analysis - especially for normal and binomial models - is sufficiently popular among users, so that a robust extension would not be considered too complicated. Having said that, however, extending the methodology to more complex models and considering more sophisticated classes of priors is of secure interest, at least from a methodological point of view. We hope to elaborate on these topics in the future.

REFERENCES

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

Armitage, P., Berry, G. and Matthews, J.N.S. (2002). *Statistical methods in medical research*. IV Edition. Blackwell Science.

Berger, J.O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis* (J. Kadane, ed.), Amsterdam: North-Holland.

Berger, J.O. (1990). Robust Bayesian analysis: sensitivity to the prior. *The Journal of Statistical Planning and Inference*, 25, 303-328.

Berger, J.O., and Berliner, L.M. (1986). Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics*, 14,461-486.

Berger, J.O., Rios Insua, D. and Ruggeri, F. (2000). Bayesian robustness. In *Robust Bayesian analysis* (D. Rios and F. Ruggeri, eds.). Lecture Notes in Statistics, 152. New York: Springer-Verlag.

Carlin, B.P., and Perez, M.E. (2000). Robust Bayesian analysis in medical and epidemiological settings. In *Robust Bayesian analysis* (D. Rios and F. Ruggeri, eds.). Lecture Notes in Statistics, 152. New York: Springer-Verlag.

Carlin, B.P., and Sargent, D.J. (1996). Robust Bayesian approaches for clinical trails monitoring. *Statistics in Medicine*, 15, 1093-1106.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, 10, 237-308.

Clarke, B.S., and Yuan, A. (2005). A closed form expression for Bayesian sample sizes. *Annals of Statistics*, to appear.

DasGupta, A. (1996). Review of optimal Bayes designs. In *Design and Analysis of Experiments. Handbook of Statistics* 13, 1099-1147.

DasGupta, A. and Mukhopadhyay, S. (1994). Uniform and subuniform posterior robustness: the sample size problem. *The Journal of Statistical Planning and Inference*, 40, 189-200.

De Santis, F. (2004). Using historical data for Bayesian sample size determination. *Rapporto Tecnico* n.11. Dipartimento di Statistica, Probabilità e Statistiche Applicate. Università di Roma "La Sapienza".

De Santis, F. (2005). Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association*, to appear.

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*, M. Di Bacco et al. editors. Kluwer Academic Publishers, Norwell, MA, USA.

Fayers, P.M., Cuschieri, A., Fielding, J., Craven, J., Uscinska, B., Freedman, L.S. (2000). Sample size calculation for clinical trials: the impact of clinicians beliefs. *British Journal of Cancer*, 82, 213-9.

Greenhouse, J.B. and Wasserman, L. (1995). Robust Bayesian methods for monitoring clinical trials. *Statistics in Medicine*, 14, 1379-1391.

Greenhouse, J.B. and Wasserman, L. (1996). A practical robust method for Bayesian model selection: a case study in the analysis of clinical trials (with discussion). In: *Bayesian Robustness, IMS Lecture Notes - Monograph Series* (J.O. Berger et. al., eds.), 331-342. Hayward: IMS.

Ianus, I. (2000). Approximate robust Bayesian inference with applications to sample size calculation. Ph.D thesis. Department of Statistics, Carnegie Mellon University.

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.

Julious, S.A. (2004). Sample sizes for clinical trials for normal data. *Statistics in Medicine*, 23, 1921-1986.

Kass, R.E., and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 435, 1343-1370.

Pericchi, L.R. and Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, 58, 1, 1-23.

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.

Sargent, D. J. and Carlin, B. P. (1996). Robust Bayesian design and analysis of clinical trials via prior partitioning (with discussion). In: *Bayesian Robustness, IMS Lecture Notes - Monograph Series* (J.O. Berger et. al., eds.), 331-342. Hayward: IMS.

Spiegelhalter, D.J, Abrams, K.R. and Myles, J.P. (2004). *Bayesian approaches to clinical trials and health-care evaluation.* 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. *Statistics in Medicine*, 5, 1-13.

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. *Statistical Science*, 17, n. 2, 193-208.

Wasserman, L. (1989). A robust Bayesian interpretation of likelihood regions. *Annals of Statistics*, 17, 1387-1393.

Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In *Bayesian Statistic 4* (J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith, eds). Oxford: Oxford University Press.