Robust Bayesian sample size determination in clinical trials

Pierpaolo Brutti, Fulvio De Santis and Stefania Gubbiotti [1]

Sapienza Università di Roma

Abstract

This article deals with determination of a sample size that guarantees the success of a trial. We follow a Bayesian approach and we say an experiment is *successful* if it yields a large posterior probability that an unknown parameter of interest (an unknown treatment effect or an effects-difference) is greater than a chosen threshold. In this context, a straightforward sample size criterion is to select the minimal number of observations so that the predictive probability of a successful trial is sufficiently large. In the paper we address the most typical criticism to Bayesian methods - their sensitivity to prior assumptions - by proposing a robust version of this sample size criterion. Specifically, instead of a single distribution, we consider a *class* of plausible priors for the parameter of interest. Robust sample sizes are then selected by looking at the predictive distribution of the lower bound of the posterior probability that the unknown parameter is greater than a chosen threshold. For their flexibility and mathematical tractability, we consider classes of ε-contamination priors. As specific applications we consider sample size determination for a Phase III trial.

Keywords: Analysis and design priors; Bayesian power; Bayesian robustness; conditional and predictive power; evidence; ε-contaminated priors; phase II and phase III clinical trials; sample size determination;

1 Introduction

In clinical trials - which will be in the following the reference context of applications - one wants to assess whether the efficacy of a new treatment (Phase II trials) or the effects difference between two treatments (Phase III trials) are larger than a chosen clinically relevant threshold. The underlying statistical problem is essentially testing a one-sided hypothesis for an unknown parameter. We are here interested in choosing an appropriate sample size for these experiments.

The most widely used sample size determination (SSD) method for testing an hypothesis on an unknown parameter is based on the classical power function, the probability of rejecting

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

the null hypothesis. This probability is computed with the sampling distribution of the data. Typically, one considers the power function evaluated at specific values of the unknown parameters under the alternative hypothesis (*conditional power*) and chooses the smallest sample size such that this quantity is sufficiently large [Armitage, Berry and Matthews (2003)]. Conditional power has been often criticized since it depends critically on the chosen design values, whose uncertainty is not accounted for. This local optimality is typical of standard classical designs and may lead to serious miscalculation of the sample size [Chaloner and Verdinelli (1995)]. To avoid local optimality, several authors have advocated a Bayesian look at the design problem. This approach allows one to model uncertainty on the design values of the parameters with a probability distribution. This is employed to average the sampling distribution, obtaining the *predictive distribution* for future data, used to compute the probability of rejecting the null, called *predictive power*. The resulting SSD methodology is a mixed Bayes-frequentist approach, since it takes into account prior uncertainty on the unknown parameters only for the design of the experiment whereas a standard classical test statistics is used for final inference [Spiegelhalter and Freedman (1986) and Joseph, du Berger and Belisle (1997)].

In this paper we consider a fully Bayesian approach to SSD, that: (*i*) models prior uncertainty on the design value when planning the experiment; and (*ii*) combines pre-experimental information with data for final inference. Specifically, we assume that an experiment is successful if one observes a large posterior probability that the unknown parameter is greater than a chosen threshold. The sample size can then be determined by looking at the predictive distribution of this posterior probability. More specifically, one considers suitable summaries - expectation and tail probabilities - of this predictive distribution and chooses the smallest number of units so that these quantities are sufficiently large.

The use of a specific elicited prior for posterior analysis has been always a major criticism to Bayesian inference. This is due to the high degree of subjectivism intrinsic to the selection of a specific distribution. An attempt to address this objection is represented by the *robust Bayesian approach* that: *a*) replaces a single prior with a class of distributions that gives a more flexible and realistic representation of pre-experimental knowledge; *b*) studies how posterior inference changes as the prior varies over the class. The idea is simple: if the range of posterior quantities of interest, the differences between the various priors in the class are irrelevant and then one can use the starting prior with confidence. On the contrary, if the posterior range is not small enough, robustness is a concern and refinement of prior knowledge is needed. General principles and developments of the robust Bayesian approach are discussed in Berger (1984, 1990), Berger, Rios Insua and Ruggeri (2000) and Wasserman

(1992). Applications of robust Bayesian analysis to clinical trials are in Greenhouse and Wasserman (1995 and 1996), Carlin and Perez (2000), Carlin and Sargent (1996) and Sargent and Carlin (1996).

This article applies the robust Bayesian philosophy to the preceding SSD problem. The basic goal is the introduction of robust SSD criteria which take into account deviations from an elicited base prior distribution for the unknown parameter. For this reason, we replace a single base prior with an entire class of distributions close to it. We assume that an experiment is successful if - as the prior varies in the class - the lower bound of the posterior probability that the parameter is greater than a chosen threshold is sufficiently large. *Robust sample sizes* are selected by looking at summaries of the predictive distribution of this lower bound.

Typically, robust sample sizes are larger than single-prior sample sizes. One of the goals of the present paper is to show the inflate of sample sizes determined by using specific classes of priors in the place of a single base prior. However, we are also interested in those circumstances (and classes of priors) in which single-prior sample sizes do not differ substantially from robust sample sizes. In these cases we say that single-prior sample sizes are robust with respect to the class Γ and that the standard procedure provides adequate sample sizes.

To model uncertainty on the base prior we here consider classes of ε-contaminated priors, studied for instance in Sivaganesan and Berger (1989). These are mixtures of the base prior with classes of distributions that possess some specific features. In this paper we focus on two specific classes of contaminating priors. The former is the set of all probability distribution, which is the largest contaminating class one can consider. The latter is the class of symmetric and unimodal distributions. These two classes of priors have been very popular in the literature on Bayesian robustness, both for being analytically tractable and also for giving fairly realistic representation of prior beliefs and uncertainty.

The present paper is related to the literature on Bayesian experimental design. For general reviews and for discussions on robust Bayesian design, see Chaloner and Verdinelli (1995), Wang and Gelfand (2002), Sahu and Smith (2006) and DasGupta (1996). For Bayesian SSD see also Joseph and Belisle (1997) and Clarke and Yuan (2006). Sample size determination methods for robust Bayesian analysis, closely related to those presented in this article, are proposed in DasGupta and Mukhopadhyay (1994), Ianus (2000), De Santis (2006) and Brutti and De Santis (2007).

The outline of the paper is as follows. Section 2 introduces the basic technical definitions of successful trial and the related SSD criteria. The concepts of robust-successful experiment and the corresponding SSD criteria are also formalized. Section 3 develops the main results in

the case of normal models, a common assumption in bio-medical applications. Implementation with two classes of ε-contamination priors are discussed and some examples presented. Section 4 reports final remarks.

2 Methodologies

Let θ be a real one-dimensional parameter denoting an unknown quantity of interest. Also, let Y_n be an estimator of θ, based on n observations, $\mathbf{y}_n$ its observed value and $f_n(\cdot;\theta)$ its density or probability mass function. We assume to be interested in assessing whether θ is larger than a threshold, δ. In Phase II superiority trials, for instance, θ may represent the effect of a treatment, an odds or an hazard function and δ a minimal "clinically significant effect". Similarly, in Phase III experiments, θ may denote an effects-difference, an odds or hazard ratio and δ a minimally "clinically significant difference" between two treatments. In the next section we review two SSD criteria for this problem. In Section 2.2 we give the corresponding robust versions.

2.1 Sample size determination criteria

Let π_A denote the prior probability distribution of θ. This is the *analysis prior*, that formalizes pre-experimental information and uncertainty on the unknown parameter θ. Given y_n, let $\pi_A(\theta|y_n) \propto \pi_A(\theta) \times f_n(y_n;\theta)$ be the posterior distribution of θ and $P_{\pi_A}(\cdot|y_n)$ the corresponding posterior probability measure. The experiment which yields the data will be said *successful* if the posterior probability that $\theta > \delta$ is larger than a chosen value, γ:

$$P_{\pi_A}(\theta > \delta|y_n) > \gamma, \qquad \gamma \in (0,1).$$

Before conducting the experiment, Y_n and $P_{\pi_A}(\theta > \delta|Y_n)$ are random. Using, for simplicity, the notation for the continuous case, let m_{π_D} denote the predictive density of Y_n:

$$m_{\pi_D}(y_n) = \int_\theta f_n(y_n;\theta)\pi_D(\theta)d\theta,$$

where π_D is the *design prior* for θ. This distribution models uncertainty on the design values for θ and, in general, it does not coincide with π_A. Note that, if π_D is a point-mass probability on a value θ_D, then m_{π_D} coincides with the sampling density $f_n(\cdot;\theta_D)$, the distribution commonly used for the standard classical sample size approach. Generally, most of Bayesian SSD criteria use the same prior distribution for computing both the posterior and the predictive distribution: see, for instance, Lindley (1997) and Raiffa and Schlaifer (2000). However, several authors have argued that two distinct priors should be used: one

prior to model *uncertainty* on the design value of the parameter and to obtain the predictive distribution; and another prior to model pre-experimental *information*, often represented by historical data, and to obtain the posterior. In this regard see, for instance, Etzioni and Kadane (1993), O'Hagan and Stevens (2001), Wang and Gelfand (2002), Sahu and Smith (2006) and De Santis (2006).

Let us consider an extreme but illustrative example for realizing why it is necessary to involve two priors. Suppose we want to design a Phase II experiment under the assumption of a large treatment effect, θ_D, but that we are uncertain on what the level of θ_D may be. In this case we use a prior π_D centered on a large guessed value θ_D, with a certain variance. At the same time, suppose that a regulatory agency requires the data from the trial to be summarized without introducing any pre-experimental bias. In this case one can use a noninformative analysis prior, which does not alter information from the data.

With no loss in generality, in designing the experiment let us assume that θ is larger than δ. This condition is formalized by choosing π_D centered on a suitably "large" value θ_D. We expect that, as n increases, the predictive distribution of $P_{\pi_A}(\theta > \delta|Y_n)$ tends to concentrate around larger and larger values. Hence, we choose the smallest n so that a suitable summary of this predictive distribution is sufficiently large. According to the summaries that we employ, different SSD criteria can be defined. We consider the following two specific cases.

1. Predictive Expectation Criterion. Let

$$e_n = \mathbb{E}_{m_{\pi_D}}[P_{\pi_A}(\theta > \delta|Y_n)]$$

be the expected value of the random posterior probability $P_{\pi_A}(\theta > \delta|Y_n)$ with respect to m_{π_D}, the predictive distribution of the data. The chosen sample size is then

$$n_e^* = \min\{n \in I\!N : \ e_n > \eta\}. \tag{1}$$

This approach is called *effect-size criterion* by Wang and Gelfand (2002).

2. Predictive Probability Criterion. Consider the predictive probability of obtaining a successful experiment:

$$p_n = \mathbb{P}_{m_{\pi_D}}[S_n] = \int_{S_n} m_{\pi_D}(y_n)dy_n,$$

where $\mathbb{P}_{m_{\pi_D}}$ is the predictive probability measure associated to m_{π_D} and S_n the subset of the sample space containing all the samples which yield a successful experiment at level γ:

$$S_n = \{y_n : \ P_{\pi_A}(\theta > \delta|y_n) > \gamma\}.$$

The chosen sample size is the smallest number of observations such that p_n is larger than a chosen threshold, η:

$$n_p^* = \min\{n \in I\!N : \ p_n > \eta\}, \qquad \eta \in (0,1). \tag{2}$$

Note that p_n is also known as *Bayesian power* (Spiegelhalter et al. (2004)).

Criterion 1 guarantees only an average control on the predictive distribution of $P_{\pi_A}(\theta > \delta|y_n)$. Criterion 2 controls also its sampling variability.

In the following we will consider examples of both methods. See De Santis (2006) for discussion on the different sensitivity to priors of expectation criteria with respect to criteria based on tail areas. Two remarks are now in order. The first is that the two-priors scheme is a general framework incorporating, as special cases, other approaches, such as the hybrid Bayes-likelihood method (π_A noninformative, π_D proper) or even classical SSD (π_A noninformative, π_D a point-mass prior on a design value θ_D). See Spiegelhalter, Abrams and Myles (2004) for discussion on conditional versus predictive versions of classical and Bayesian power. The second remark is more technical. Generally, at least in standard models, a noninformative analysis prior leads to a proper posterior. Conversely, a noninformative improper design prior cannot be employed since the corresponding marginal distribution of the data, m_{π_D}, is undetermined. See, for instance, De Santis (2007) for discussion on this point.

2.2 Robust sample size determination criteria

Suppose now that, instead of a single analysis prior distribution, we are only able to elicit a class of distributions Γ_A. Specifically, assume we single out a prior π_0 that quantifies pre-trial information on θ, but that we are not completely confident in it. We replace π_0 with a class of distributions "close" to it. The trial is considered *robust-successful* if, for any prior in Γ_A, the posterior probability that θ is greater than δ is larger than γ or, equivalently, if

$$\inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta|y_n) > \gamma, \qquad \gamma \in (0,1).$$

The robust version of SSD criteria 1 and 2 defined above are obtained by simply replacing $P_{\pi_A}(\theta > \delta|y_n)$ with $\inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta|y_n)$ in (1) and (2).

1. Robust Predictive Expectation Criterion. Let

$$e_n^r = \mathbb{E}_{m_{\pi_D}}[\inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta|Y_n)].$$

The, for $\eta \in (0,1)$, the selected sample size is

$$n_{e,r}^* = \min\{n \in I\!N : \ e_n^r > \eta\}. \tag{3}$$

This is the *robust effect-size criterion.*

2. Robust Predictive Probability Criterion. For given a $\eta \in (0,1)$, the selected sample size is

$$n^*_{p,r} = \min\left\{n \in I\!N : \ p^r_n > \eta\right\}, \tag{4}$$

where

$$p^r_n = \mathbb{P}_{m_{\pi_D}}[S^r_n] = \int_{S^r_n} m_{\pi_D}(y_n)dy_n$$

denotes *robust predictive power* and

$$S^r_n = \left\{y_n : \ \inf_{\pi_A \in \Gamma_A} P_{\pi_A}(\theta > \delta | y_n) > \gamma\right\},$$

i.e. the subset of the sample space whose elements, for each prior in Γ_A, yields a posterior probability that $\theta > \delta$ larger than γ.

Generally, the consequence of replacing π_A with Γ_A (which we assume to contain π_A), is that, for any given δ, η and γ, the robust sample size is larger than the single-prior sample size. Similarly, for any two classes of priors Γ_A and Γ'_A such that $\Gamma_A \subset \Gamma_{A'}$, optimal sample sizes determined with the latter class are larger than those obtained with the former. Numerical examples will be discussed in Section 3.

2.3 Robust sample size determination with ε-contamination classes

An ε-contamination class, studied for instance in Berger and Berliner (1986) and Sivaganesan and Berger (1989), is defined as follows:

$$\Gamma_\varepsilon = \{\pi_A : \ \pi_A(\theta) = (1-\varepsilon)\pi_0(\theta) + \varepsilon q(\theta); \ q \in Q\},$$

where π_0 is the base prior, $\varepsilon \in (0,1)$ is the degree of contamination we consider for this distribution and q is the contaminant prior, that varies in a given class Q. According to the choice of Q, we have different ε-contamination classes. Among the many available, we consider the following two classes for Q:

$Q_{US} = \{$all the unimodal and symmetric distributions with the same mode μ_0 as that of $\pi_0\}$;

and

$$Q_{All} = \{\text{all the distributions}\}.$$

The corresponding ε-contamination classes will be denoted respectively as Γ_{US} and Γ_{All}. The class Q_{All} is appealing for its analytical tractability but it contains many more priors than we would often consider plausible in practice. We will see in the following sections that this

fact determines very large sample sizes even for small amounts of contamination. The class Q_{US} is still analytically feasible but it restricts considerably the set of possible contaminant distributions compared to Q_{All}. Sivaganesan and Berger (1989) provides the expressions for lower and upper bounds of the posterior probability of a set H as the prior varies in Γ_{US} and Γ_{All}. Note that, for any y_n,

$$\Gamma_{US} \subset \Gamma_{All} \qquad \Rightarrow \qquad \inf_{\pi_A \in \Gamma_{US}} P_{\pi_A}(H|y_n) \geq \inf_{\pi_A \in \Gamma_{All}} P_{\pi_A}(H|y_n).$$

Hence, optimal sample sizes computed using Γ_{US} are smaller than those determined with Γ_{All}.

For both the classes considered here, closed-form expressions for e_n^r and p_n^r cannot be determined and we will resort to standard Monte Carlo approximations: we draw a large number of samples from the predictive distribution of the data, m_{π_D}, and for each generated value $\tilde{y}_n(j)$ we compute $\inf_{\pi_A \in \Gamma_\varepsilon} P_{\pi_A}[H|\tilde{y}_n(j)]$. These quantities are obtained exploiting the results by Sivaganesan and Berger (1989) (see Appendix A.1) and are then used to determine numerical approximations for e_n^r and p_n^r.

3 Results with normal likelihoods

Assume now that data relevant to θ are summarized by a statistic Y_n with (at least approximately) normal distribution of parameters $(\theta, \sigma^2/n)$. In Phase II clinical trials, for instance, θ may denote a treatment effect, n the number of individuals assigned to the treatment, Y_n the sampling mean of experimental outcomes normally distributed with expectation θ and variance σ^2. However, the same basic model provides an approximation that can be used, for instance, for binary data - with θ denoting a log-odds - and for survival data - with θ denoting a log-hazard function - (see Spiegelhalter et al., 2004, Sections 2.4.1 and 2.4.2). Note that, if σ^2 is unknown, a full Bayesian analysis with a proper prior (such as, for example, inverted-gamma priors) can be performed but, for brevity, we will omit this case and will rely on the simplifying assumption that σ^2 is known. See Spiegelhalter and al. (2004) for discussion on this point.

For computational simplicity, we assume $\pi_0(\theta) = N(\theta|\theta_0, \sigma^2/n_0)$, where $\mathrm{N}(\cdot|a, b)$ denotes the density function of a normal random variable of mean a and variance b and n_0 is the so-called "prior sample size". It follows that

$$P_{\pi_0}(\theta > \delta|y_n) = 1 - \Phi\left(\frac{\delta - \theta_{\pi_0}(y_n)}{\sigma_{\pi_0}}\right) \tag{5}$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal random variable, $\theta_{\pi_0}(y_n) = (n_0\theta_0 + ny_n)(n_0 + n)^{-1}$ and $\sigma_{\pi_0} = \sigma^2(n + n_0)^{-1/2}$ are the posterior expectation and standard deviation of θ under the base prior π_0.

As design prior, we consider $\pi_D(\theta) = N(\theta|\theta_D, \sigma^2/n_D)$, where θ_D is the design value around which we spread the probability mass, with precision determined by n_D. In this case, $m_{\pi_D}(y_n) = N(y_n|\theta_D, \sigma^2(1/n + 1/n_D))$. Note that for $n_D \to +\infty$, the predictive distribution m_{π_D} tends to the sampling distribution $f_n(\cdot;\theta_D)$, the density used in standard classical sample size choice.

The use of conjugate design and analysis priors yields a closed-form expressions for p_n. Under the above assumptions, for a given $\eta \in (0,1)$, we have

$$p_n = \Phi\left(\frac{\xi - \theta_D}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n_D}}}\right), \tag{6}$$

where $\xi = (n_0 + n)(\delta - z_{1-\eta}\sigma_{\pi_0})/n - n_0\theta_0/n$ and where z_α denotes the α-level percentile of the standard normal distribution. See Spiegelhalter et al. (2004) for the $\pi_0 = \pi_D$ special case. Implementation of robust SSD methods with ε-contamination classes Γ_{US} and Γ_{All} is presented in Appendix A.2.

3.1 Examples: robust SSD for inference on a log-hazard ratio

Suppose we want to design a trial for inference on a log-hazard ratio, θ, using the statistic $Y_n = 4L_n/n$, where L_n is the standard log-rank statistic and n the total number of events (deaths) that will be observed in the trial. It is assumed that Y_n is approximatively normal with parameters $(\theta, \sigma^2/n)$, with $\sigma^2 = 4$ (see Spiegelhalter et al., 2004, Section 2.4.2, for details). We revisit an example from Spiegelhalter et al. (2004, Examples 2.6 and 6.2) in which a balanced trial is designed to have 80% of classical power to detect a log-hazard ratio $\theta_D = 0.56$, equivalent to a raise of 5-year survival from 20% to 40% in favor of the new treatment. The Authors consider a design prior, centered on the guessed value θ_D and with 0.05 probability that θ is less than zero (old treatment better than new treatment). This results in a design-prior sample size $n_D = 34.5$ and, overall, in a design density which represents optimism towards the new treatment. The prior is then employed to average the classical power curve and to obtain an hybrid classical-Bayes power to be compared with the standard procedure. This is equivalent to using a noninformative analysis prior for θ.

We here extend Spiegelhalter et al.'s example with the introduction of a base analysis prior π_0 and with the robust analysis illustrated in previous sections. Specifically, as base analysis prior π_0, we start considering a normal density centered on $\theta_0 = 0$, expressing equivalence between old and new treatment, and variance such that the probability that θ is greater than θ_D is equal to a chosen value α. This choice yields an analysis prior sample size n_0 equal to $(2z_{1-\alpha}/\theta_D)^2$. Note that, the smaller the values of α and of $|\theta_D|$, the more sceptical the resulting base prior. Of course an equivalent way to define a sceptical base prior is to fix

θ_D and then set θ_0 to a value close to 0 and smaller than θ_D. For instance if the guessed value is $\theta_D = 0.56$, we can choose $\alpha = 0.2$, so that, on the one hand we assign low chance to the values of the parameter greater than θ_D, and on the other we are still allowing for a relatively high uncertainty, corresponding to a low value of the prior sample size, namely $n_0 = 9$. The analysis base prior is therefore less informative than the design prior. In addition, we assume a minimal clinically significant difference $\delta = 0.1$, corresponding to a raise in the survival rate from 20% to 23.3%.

The contour plot in Figure 1 represents the lower bound of e_n^r for Γ_{All} as the sample size n and the contamination parameter ε vary. We notice that, even for low levels of contamination, the sample size required to reach $\eta = 0.8$ ($n^*_{e,r} = 124$, for $\varepsilon = 0.1$), is substantially larger than the standard optimal sample size ($n^*_e = 56$). Nevertheless, if we are willing to slightly reduce η, for example to values around 0.7, we are able to achieve significantly smaller sample sizes ($n^*_{e,r} \sim 60$) even for a moderate amount of contamination ($\varepsilon \sim 0.2$). The optimal sample sizes listed above are clearly unreasonable in many practical

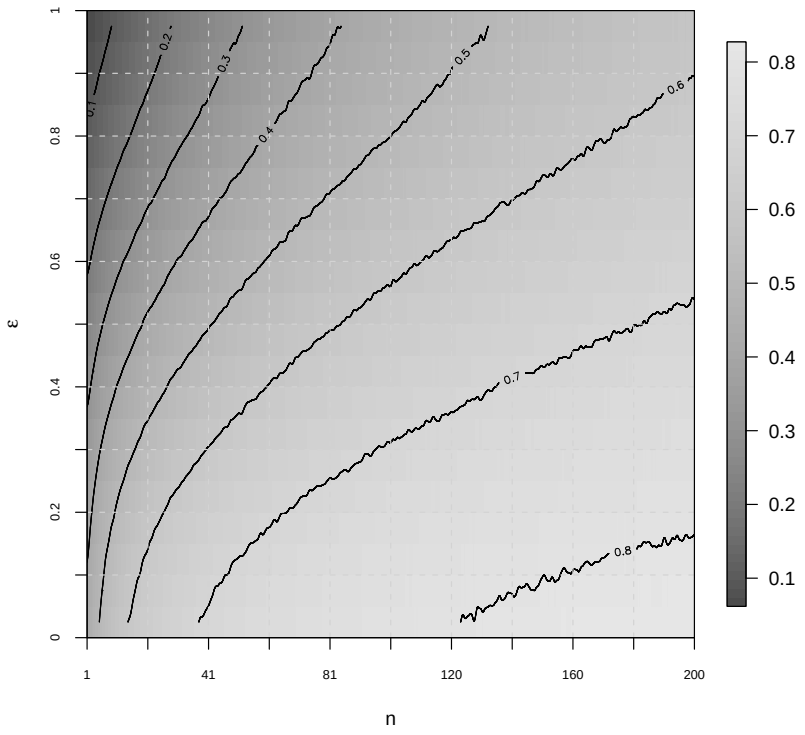

Figure 1: Contour plot of e_n^r for Γ_{All} as the sample size n and the contamination parameter ε vary, assuming: $\sigma^2 = 4$, $\theta_0 = 0$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$.

situations. This is a consequence of the content itself of the contamination class which contains many undesirable distributions such as point masses that are far way from the base

prior π_0.

As already discussed in Section 2.3, a plausible alternative contamination class is Γ_{US}. In Table 1 we summarize the standard and robust optimal sample sizes computed for both classes Γ_{All} and Γ_{US}, and for different levels of contamination. Focusing on the rows related to Γ_{US} the overall impression is that the optimal sample sizes we obtain are extremely stable with respect to the contamination level, when compared to what happens under the class Γ_{All}. The same conclusions can be drawn by looking at Figure 2. In fact, as shown in the right panel of this graph, the distance between the two extrema related to Γ_{US} is actually negligible even for values of ε approaching 1. In order to observe a wider distance, we

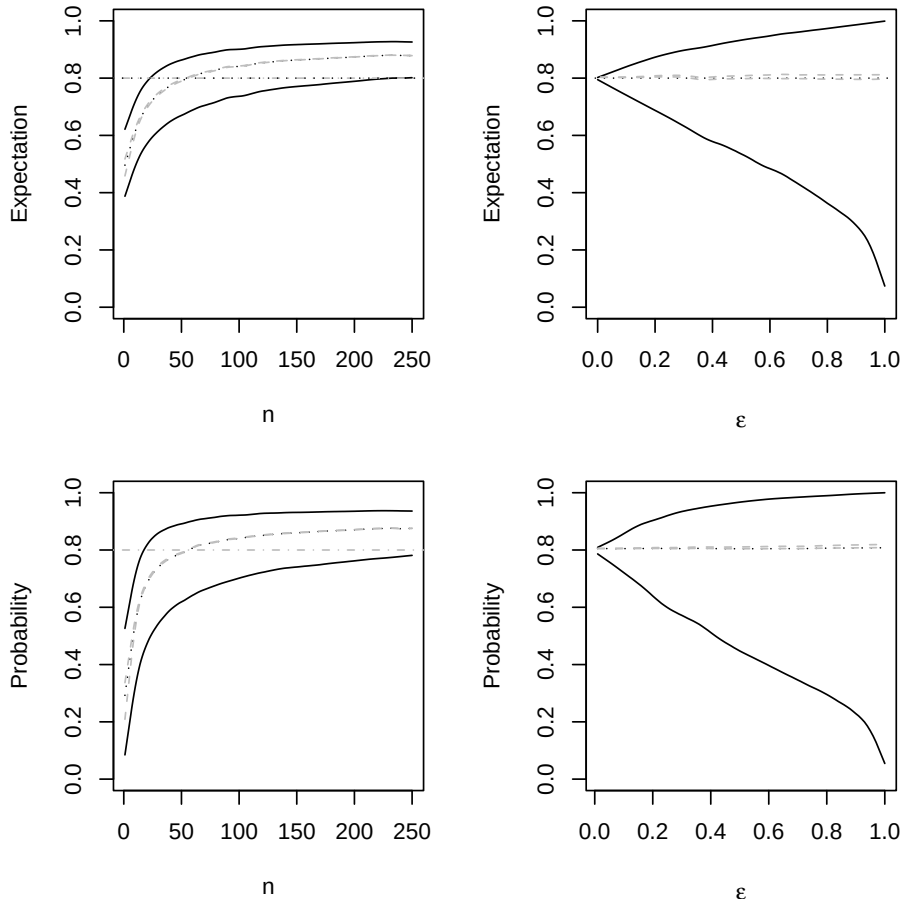

Figure 2: e_n^r (first row) and p_n^r (second row) for Γ_{All} (solid line) and Γ_{US} (dashed line) as the sample size n (first column, with $\varepsilon = 0.2$) and the contamination parameter ε (second column, with $n = n_e^* = n_p^* = 56$) vary, assuming: $\sigma^2 = 4$, $\theta_0 = 0$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$. The horizontal reference line is set to $\eta = 0.8$.

can force the two priors π_0 and π_D to express radically opposite beliefs. For example, we might center the analysis base prior on $\theta_0 = -1.6$, expressing a very pessimistic opinion on the experimental treatment and, conversely, the enthusiastic design prior on $\theta_D = 1.6$,

corresponding to a hazard ratio equals to 5 in favor of the new treatment. In this extreme situation depicted in Figure 3, the predictive expectation criterion based on e_n^r leads to more cautious conclusions than the standard criterion e_n.

		Expectation		Probability	
Class	ε	$\theta_0 = 0$	$\theta_0 = 0.29$	$\theta_0 = 0$	$\theta_0 = 0.29$
All	0.1	124	100	162	128
	0.2	236	209	359	329
	0.3	366	338	520	499
US	0.1	56	40	56	36
	0.2	57	40	57	37
	0.3	57	41	57	38
Standard	0.0	56	39	56	36

Table 1: Optimal sample sizes $n^*_{e,r}$ and $n^*_{p,r}$ for Γ_{All} and Γ_{US} and 3 different levels of contamination ($\varepsilon \in \{0.1, 0.2, 0.3\}$), assuming: $\sigma^2 = 4$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$, $\eta = 0.8$, $\gamma = 0.6$ and two different base analysis priors π_0, namely a sceptical one ($\theta_0 = 0$) and a enthusiastic one ($\theta_0 = 0.29$). The line labelled Standard contains the non–robust optimal sample sizes n^*_e and n^*_p (associated to $\varepsilon \equiv 0$).

Moving to the predictive probability criterion, the right side of Table 1 summarizes our findings. As for Γ_{US}, we reach similar conclusions to those just obtained for the expectation criterion, whereas the optimal sample sizes induced by Γ_{All} are even larger than before because of the higher sensitivity of this criterion to the presence of extreme distributions in the contamination class. All these results are strongly influenced by the value of the parameter γ. In our case the chosen $\gamma = 0.6$ leads to optimal sample sizes comparable to those selected by the e_n^r. Of course increasing γ would result in larger and larger values of the optimal sample size.

As mentioned above, once we fix the design mean θ_D to 0.56, shifting the mean of the base prior from $\theta_0 = 0$ to an intermediate value between 0 and 0.56, for example to $\theta_0 = 0.29$, results in a more optimistic opinion about the experimental treatment. Consequently the optimal sample sizes associated to $\theta = 0.29$ in Table 1 are uniformly smaller than those obtained using the sceptical base prior. It is quite interesting to notice that in the extreme case in which the analysis and the design priors are coincident we observe that e_n, p_n and their robust versions tend to be flat for large enough values of n (see Figure 4). This can be explained by the impossibility of keeping the same interpretation for the design prior: in this setting, the reference value θ_D does not express optimism anymore with respect to the

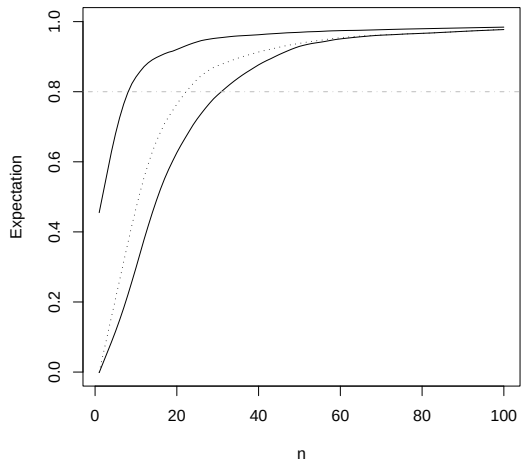

Figure 3: e_n^r for Γ_{US} as n varies, assuming: $\varepsilon = 0.2$, $\sigma^2 = 4$, $\theta_0 = -1.6$, $n_0 = 9$, $\theta_D = 1.6$, $n_D = 9$, $\delta = 0.1$. The horizontal reference line is set to $\eta = 0.8$, whereas the dotted line corresponds to the standard (non–robust) criterion e_n.

beliefs represented by the base analysis prior.

Finally we focus on Figure 5, where we compare the proposed robust power curves with their Bayesian and classical counterparts. As in Spiegelhalter et al. (2004), choosing an enthusiastic prior ($\theta_0 = 0.56$), the Bayesian power curve is substantially higher than the classical one because, in their terms, the prior gives a "head start". On the contrary, the effect of contamination with Γ_{All} on the robust power curve gives rise to lower power as the contamination level ε increases. As an example, choosing $\varepsilon = 0.2$ results in a robust power curve significatively lower than the classical one (see Figure 5).

4 Discussion

The use of robust techniques in a Bayesian framework allows to address the critical dependence of the inferential conclusions on the specification of a prior distribution. The present paper deals with this problem in the pre-experimental context, when the size of a trial has to be selected. The main message of the paper is that, in the presence of uncertainty in prior specification, the sample size should be adequately larger than it is in the presence of more refined knowledge. The goal is avoiding sample sizes smaller than necessary, that would imply a low predictive probability of success for the trial.

We have given a robust Bayesian look at the sample size determination problem in clinical trials and proposed some criteria which extend the more traditional approaches based on the power of a test. In order to take into account uncertainty on the base prior, we have

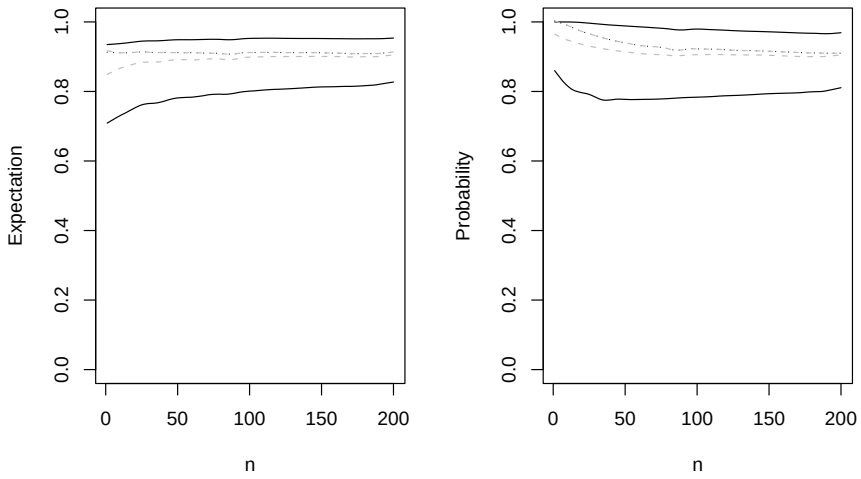

Figure 4: e_n^r (left) and p_n^r (right) for Γ_{All} (solid line) and Γ_{US} (dashed line), assuming: $\varepsilon = 0.2$, $\sigma^2 = 4$, $\theta_0 = 0.56$, $n_0 = 34.5$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$

replaced it with an entire class of priors and considered the resulting robust sample sizes. In the context of normal models, we have showed examples in which sample sizes selected using the base prior are very close to robust sample sizes, obtained using the class of symmetric and unimodal distribution. We have also seen that relevant discrepancies between single-prior and robust sample sizes are obtained only in the presence of a dramatic difference between design and analysis priors. The robustness of the standard Bayesian procedure observed in the examples of Section 3 is interesting whenever the class Γ_{US} is a fairly reasonable representation of prior beliefs on θ. Basically, we now know that using sample sizes based on a normal base prior are still adequate under contamination, as long as the contaminating priors respect the constraints of symmetry and unimodality.

We have also shown that, in the same examples and even for modest contamination levels, using Γ_{All} implies samples sizes in general quite larger than those found with the base prior π_0. One can object that the class Γ_{All} is "too big", containing unreasonable prior distributions for the parameter. But we have used this class as a "worst case": at chosen ε levels, robust sample sizes selected using Γ_{All} automatically satisfy SSD criteria for any other contamination class. Of course, one can consider refinements of this class and select sample sizes appropriate to the available prior knowledge. One possibility is the class of unimodal distribution (Sivaganesan and Berger, 1989). Preliminary numerical studies in the context of the examples of Section 3 shows robust sample sizes obtained with this class are close to those found using Γ_{US}. We will elaborate more on this in future research, as well as considering other classes of prior distributions. We also hope to discuss on sensitivity to the model (deviation from normality) and applications of the proposed robust criteria to other

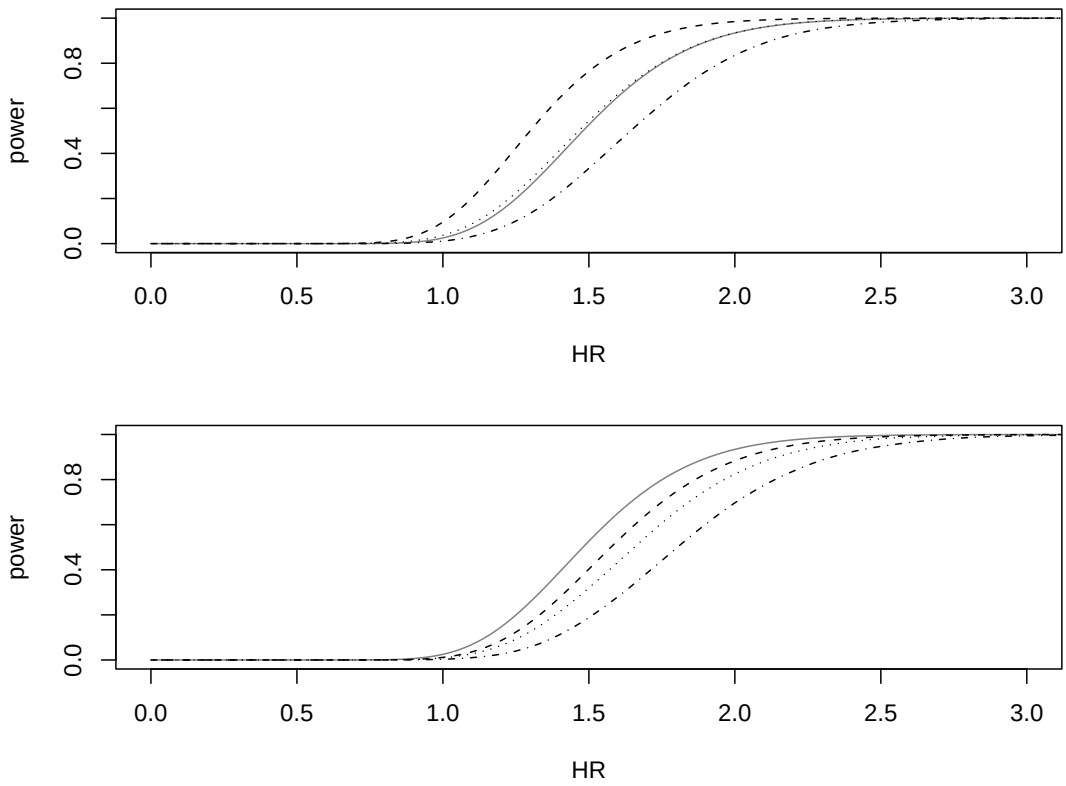

Figure 5: Power functions: classical (solid line), Bayesian (dotted line), and robust for Γ_{All} with $\varepsilon = 0.05$ (dashed line) and $\varepsilon = 0.2$ (dashed-dotted line), assuming respectively an enthusiastic prior, centered on $\theta_0 = 0.56$ (top) and a sceptical prior, centered on $\theta_0 = 0$ (bottom).

parametric models (binary end-points).

We have discussed the necessity of a suitable trade-off between the level of contamination and the class Q, on the one hand, and the chosen thresholds δ, η and γ, on the other. The idea is simply that, in fixing the goals of an experiment, one should take into account the degree of uncertainty on the prior, represented by the class Q and by ε: a large degree of uncertainty on the prior implies in general unrealistic large sample sizes if the goal of the trial are too ambitious (large δ, η and γ). The message is here that the sample size problem is much more problematic than it is typically perceived in that it requires accurate modelling of both goals of the trials and available uncertainty and information.

APPENDIX

A.1. Bounds of the posterior probability of a set with symmetric-unimodal and arbitrary contaminations.

Assume that the class of contaminating distributions is Q_{US}. Sivaganesan and Berger (1989) show that, for any set H,

$$\inf_{\pi_A \in \Gamma_{US}} P_{\pi_A}(H|y_n) = \inf_{z \geq 0} \frac{a_0 + K_1(z)}{a + K_2(z)} \qquad \text{and} \qquad \sup_{\pi_A \in \Gamma_{US}} P_{\pi_A}(H|y_n) = \sup_{z \geq 0} \frac{a_0 + K_1(z)}{a + K_2(z)}, \quad (7)$$

where

$$a_0 = a P_{\pi_0}(H|y_n), \qquad a = \frac{1-\varepsilon}{\varepsilon} m_{\pi_0}(y_n), \quad (8)$$

$m_{\pi_0}(y_n)$ and $P_{\pi_0}(H|y_n)$ are respectively the marginal density of the data and the posterior probability of H, both computed with the base prior π_0;

$$K_1(z) = \begin{cases} \frac{1}{2z} \int_{\theta_0 - z}^{\theta_0 + z} I_H(\theta) f_n(y_n;\theta) d\theta & z > 0 \\ I_H(\theta_0) f_n(y_n;\theta_0) & z = 0 \end{cases}$$

with $I_H(\cdot)$ denoting the indicator function of the set H; and where

$$K_2(z) = \frac{1}{2z} \int_{\theta_0 - z}^{\theta_0 + z} f_n(y_n;\theta) d\theta.$$

For arbitrary contaminations, Berger and Berliner (1986) show that

$$\inf_{\pi_A \in \Gamma_{All}} P_{\pi_A}(H|y_n) = \frac{a_0}{a + L_{H^c}} \qquad \text{and} \qquad \sup_{\pi_A \in \Gamma_{All}} P_{\pi_A}(H|y_n) = \frac{a_0 + L_H}{a + L_H},$$

where a_0 and a are given in (8), and where $L_H = \sup_{\theta \in H} f_n(y_n;\theta)$.

A.2. Bounds of the posterior probability of a set for normal models.

Let

$$H = \{\theta : \theta > \delta\}.$$

Under normality assumptions, the explicit expressions of a_0 and a in (8) can be obtained by noting that

$$m_{\pi_0}(y_n) = N\left(y_n | \mu_0, \sigma^2 \left(\frac{1}{n} + \frac{1}{n_0}\right)\right)$$

and by using the expression of $P_{\pi_0}(H|y_n)$ given in (5). For unimodal-symmetric contaminations, bounds of $P_{\pi_A}(H|y_n)$ are obtained from (7) with

$$K_2(z) = \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 - z - y_n)\right)\right].$$

It can also be checked that for $z > 0$

$$K_1(z) = \begin{cases} K_{1,a}(z) & \theta_0 < \delta \\ K_{1,b}(z) & \theta_0 = \delta \\ K_{1,c}(z) & \theta_0 > \delta \end{cases},$$

where

$$K_{1,a}(z) = \begin{cases} 0 & z \leq \delta - \theta_0 \\ \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\delta - y_n)\right)\right] & z > \delta - \theta_0 \end{cases},$$

$$K_{1,b}(z) = \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\delta + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\delta - y_n)\right)\right]$$

and

$$K_{1,c}(z) = \begin{cases} \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 - z - y_n)\right)\right] & z \leq \theta_0 - \delta \\ \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\delta - y_n)\right)\right] & z > \theta_0 - \delta \end{cases}.$$

Using the class Γ_{All}, bounds for $P_{\pi_A}(H|y_n)$ are determined noting that, under the above normality assumptions,

$$L_H = \phi\left(\frac{\sqrt{n}(\delta - y_n)}{\sigma}\right) I_{(-\infty,\delta)}(y_n) + \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} I_{(\delta,+\infty)}(y_n)$$

and

$$L_{H^c} = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} I_{(-\infty,\delta)}(y_n) + \phi\left(\frac{\sqrt{n}(\delta - y_n)}{\sigma}\right) I_{(\delta,+\infty)}(y_n),$$

where $\phi(\cdot)$ is the density function of a standard normal random variable.

REFERENCES

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.

Brutti, P. and De Santis, F. (2007). Avoiding the Range of Equivalence in Clinical Trials: Robust Bayesian Sample Size Determination for Credible Intervals. *The Journal of Statistical Planning and Inference*. To appear.

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. (2006). A closed form expression for Bayesian sample sizes. *Annals of Statistics*, 34, n. 3, 1293-1330.

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. (2006). Sample size determination for robust Bayesian analysis. *Journal of the American Statistical Association*, 101, n. 473, 278-291.

De Santis F. (2007). Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society*, Ser. A, 170, 1, 95-113.

Etzioni, R., and Kadane, J.B. (1993). Optimal experimental design for anothers analysis. *Journal of the American Statistical Association*, 88, n. 424, 1404-1411.

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.

Lindley, D.V. (1997). The choice of sample size. *The Statistician*, 46, 129-138.

O'Hagan, A., and Stevens, J.W. (2001). Bayesian assessment of sample size for clinical trials for cost effectiveness. *Medical Decision Making*, 21, 219-230.

Raiffa H. and Schlaifer R. (2000) *Applied Statistical Decision Theory*, Wiley.

Sahu, S. K. and Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications. *Journal of the Royal Statistical Society*, Ser. A, 169, 235-253.

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.

Sivaganesan, S. and Berger, J.O. (1989). Ranges of posterior measures for priors with unimodal contaminations. *Annals of Statistics*, 17, 868-889.

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. (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.