SAPIENZA
UNIVERSITÀ DI ROMA

Dottorato di Ricerca in Statistica Metodologica

Tesi di Dottorato XXI Ciclo – 2005/2008

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Bayesian Methods for Sample Size Determination and their use in Clinical Trials

Stefania Gubbiotti

thanks to my penguins..

Al termine di questo lavoro vorrei ringraziare tutti coloro che hanno accompagnato il mio percorso da dottoranda. L'elenco non può che partire dal coordinatore Fulvio De Santis, ma non certo per formalità... piuttosto per la sua pazienza, il suo supporto costante e per tutte le opportunità di crescita che mi ha offerto sul piano scientifico e soprattutto umano. Ma l'elenco inizia qui solo grazie al mio *mentore* Pierpaolo Brutti senza il quale la storia sarebbe stata *maledettamente* diversa... E poi per l'amicizia e la condivisione di mille cose che vanno al di là del contenuto di una tesi ringrazio Marco Perone Pacifico, Isa Verdinelli e Luca Tardella che, non a caso, vanno a finire insieme a tutti gli altri *ragazzi*: Serena, Maria Brigida e Alessia con cui ho mosso i primi passi, insieme a Claudia, Antonello, Francesca e Daria, i *veterani*, fino ai veri e propri *antenati*, Alessio, Paolo, Valeria, e poi alla schiera dei *piccoli*, Valentina, Antonietta, Bianca, Mirko, Alessandra e ovviamente Federico.

Grazie anche a tutto il Dipartimento di Statistica, Probabilità e Statistiche Applicate e alle persone con cui ho collaborato nella mia parentesi svizzera presso la Novartis Pharma di Basilea, Beat Neuenschwander, Amy Racine e Michael Branson.

Infine grazie a chi mi ha sopportato anche fuori dall'Università, i miei genitori, i miei amici, Giulia e, più di tutti, Andrea.

Contents

List of Figures

List of Tables

Introduction

Sample Size Determination (SSD) – that is the choice of the optimal number of observations to be enrolled in a study in order to guarantee good quality inference – is one of the crucial aspects of experimental design. In this thesis we essentially refer to the context of clinical trials, both for the terminology and for the applications, although the proposed methodology can be applied to a more general experimental setting. More specifically, we focus on Phase II clinical trials that are aimed at evaluating new biomedical procedures in terms of efficacy and/or safety. Another interesting setting is the one of Phase III studies in which two alternative treatments are to be compared. Clinical trials constitute a broadly accepted standard framework to develop and regulate progresses in biomedical sciences and they also provide an ideal context for the implementation of innovative statistical techniques. From the SSD point of view, the main objective of a clinical trial is to recruit the minimum number of patients that guarantees to obtain conclusive inferential results with high probability. At the same time, in planning a trial one needs both to satisfy budget constraints and to care about ethical implications, related to patients' health (see Julious (2004)).

According to the classical perspective, the optimal sample size is calculated using formulae based on the power of a test or on the width of a confidence interval (see Armitage et al. (2002)). In general, frequentist procedures rely on the computation of probabilities of certain events with respect to the sampling distribution. Given that the latter depends on the unknown parameter, it is then necessary to prefix a guess value for the parameter of the assumed statistical model. This value, also called design value, has a heavy impact on the SSD criteria that finally turn out to be only locally optimal.

This is one of the motivations that encourages us to consider a Bayesian approach, that allows us to model initial uncertainty on the design parameters through a prior probability distribution. For instance, De Santis & Perone Pacifico (2004) highlight that assigning a probability distribution to the unknown design quantity allows one to compare alternative scenarios and to avoid local optimality. Furthermore, while in the frequentist approach one ignores pre-experimental information on the phenomenon of interest – for instance derived from historical studies or from subjective opinions of experts – Bayesian methods provide a rigorous framework to formalize and incorporate this information in inferential analysis. In the specific case of SSD, this potentially yields a reduction of the required number of observations to reach the prefixed objectives.

Hence, in this work we introduce suitable SSD criteria, based on specific summaries of the predictive distribution of a chosen posterior quantity of interest. We follow here the so-called *two–priors approach*. It establishes that it is possible to specify two distinct prior distributions: on the one hand the *design prior* models initial uncertainty on the parameter, on the other the *analysis prior* allows one to take into account pre-experimental information in the preposterior analysis. This topic is discussed in Chapter 1; for further details see for example Tsutakawa (1972), Etzioni & Kadane (1993), Wang & Gelfand (2002), De Santis (2006).

In Chapter 2 we highlight that, given particular choices of the posterior quantity of interest and of the predictive summary, power-based methods for SSD can be thought as a special case of the predictive Bayesian approach. This interpretation is particularly appealing in that it involves the most widely used methods in standard applications. Moreover, it allows a generalization of the notion of power function. To this end, first of all we show how the classical power – that in the following we name *Conditional Frequentist Power function* – does not take into account: (a) uncertainty on the design value used for the unknown parameter to compute the power; (b) pre-experimental information on the unknown parameter, provided, for instance, by previous clinical studies or by subjective opinions of experts. Conversely, by taking into account (a) or (b) or both, several extensions of the power function are proposed: *Predictive Frequentist Power function*, *Conditional* and *Predictive Bayesian Power functions*. We review these methods, their relationships with the standard approach and implications on sample size determination and we discuss an application with regard to the normal model (see Gubbiotti & De Santis (2008)). Finally, this leads us to notice that Predictive Bayesian Power can be interpreted as a generalized power function, including the others as special cases.

In the second part of the thesis, the general framework of Chapter 1 is extended in several directions. The first step considered in Chapter 3 is the introduction of a robust version of SSD criteria. Elicitation of a prior distribution is often criticized because of the impact that a specific prior has on preposterior analysis and on selected sample sizes. In other words, an additional amount of uncertainty should be accounted for in prior elicitation. For this reason, by replacing the single prior with a class of prior distributions, we derive a robust version of the SSD criteria. This approach actually results in larger values of the sample size, as we show in the applications with respect to the normal model and the binomial model (see Brutti et al. (2008*b*)).

As a second extension, in Chapter 4, we assume that prior information derives from several sources, for instance from distinct historical studies or from different experts opinions (see Brutti et al. (2008*a*)). Hence, we suggest to elicit a prior distribution to formalize the information relative to each of these sources and to combine these distributions through a mixture with conveniently chosen weights. This straightforward method allows us to deal with multiple sources of uncertainty: the same framework of Chapter 2 can be then used to establish predictive SSD criteria. Furthermore, we extend the use of a mixture of informative priors to the case of Sample Size Re-estimation (SSRe): assuming that during an ongoing trial, at a given time point, the first part of collected data is already available, we propose to adjust the optimal sample size chosen at the beginning of the trial, based on the interim information. This is very natural in the Bayesian context, since the information can be easily updated thanks to Bayes theorem.

In this thesis we mostly refer to superiority trials. In Chapter 5, however, we explicitly refer to equivalence trials, aimed at demonstrating no clinically significant difference between two treatments, i.e. that the competing therapies are clinically equivalent. Hence, we adapt the Bayesian SSD criteria to an equivalence study and we consider a robust version of these criteria for classes of restricted conjugate priors. Results for the normal model are provided and illustrated by examples.

Chapter 1

A predictive approach to Bayesian Sample Size Determination

1.1 Introduction and motivations

In this thesis we introduce a predictive Bayesian methodology for sample size determination in the context of clinical trials. In general, the main purpose of a clinical trial is to observe, as efficiently as possible, the minimum number of individuals allowing inferential analysis to be conclusive. However, it is clear thar the choice of the sample size is also connected to budget costraints and, above all, ethical implications. In fact, as discussed in Julious (2004), if the sample size is too large the trial could have met its objective before reaching its actual end, that is before recruiting the preplanned number of patients, so that some individuals may have unnecessarily entered the trial. On the contrary, if the trial is too small, there will be little chance of meeting the study objectives, and patients may be put through the potential trauma of a trial for no tangible benefit.

First of all, we briefly remind the reader the current classification of the main clinical trials categories, according to the FDA (see Clinicaltrials.gov (2008)):

- In **Phase I trials**, researchers test an experimental drug or treatment in a small group of people for the first time to evaluate its safety, determine a safe dosage range, and identify side effects.
- In **Phase II trials**, the purpose is to check if the treatment is effective and to further evaluate its safety.
- In **Phase III trials**, the treatment is given to a large groups of people to confirm its effectiveness, monitor side effects, compare it to commonly used

treatments, and collect information that will allow the experimental drug to be used safely.

- In **Phase IV trials**, post marketing studies delineate additional information including the drug's risks, benefits, and optimal use.

In most of the applications illustrated in this work we specifically refer to Phase II and Phase III trials. Moreover we mainly deal with superiority trials, although the presented methods can be adapted to experimental designs with different objectives; for instance in Chapter 5 we focus on equivalence trials.

According to a classical perspective, the optimal sample size is usually determined either from power calculations or from formulae based on confidence interval widths (see, for example, Armitage et al. (2002)). Both cases involve the use of the sampling distribution that depends on the unknown parameter of interest. Hence, standard frequentist procedures require initial guesses of the parameters, which implies that the resulting criteria are only locally optimal. In other words, the selected sample size can be quite sensitive to these guessed values. This drawback of standard SSD methods is discussed in details and illustrated by examples in Section 1.1.1. In order to avoid local optimality, it is possible to resort to a Bayesian approach, that specifically deals with this problem by modeling prior uncertainty on the parameter values through a prior probability distribution. As Berger (1985) said indeed, design problems are "naturally Bayesian": before the experiment is performed, the absence of data forces to address planning issues by using prior information. Bayesian methods provide a rigorous framework that allows one to incorporate either historical information derived from previous studies or subjective opinions of expert clinicians by specifying a prior distribution.

As discussed in De Santis & Perone Pacifico (2004), pre-experimental information can contribute not only to reduce the overall size of an experiment but also to efficiently allocate the experimental units, with more individuals assigned to the innovative treatment, for which it is assumed that less information is available. The Authors point out that, when the comparison of two unknown parameters representing the mean effectiveness of two treatments is of concern, using a probability distribution in order to formalize prior information on these quantities has two immediate advantages. The first is practical: assigning a prior distribution to the unknown quantities allows different plausible scenarios to be taken into consideration. Technically speaking, this allows local optimality to be avoided. Moreover, in comparing two treatments effects, the Bayesian approach allows for the use of flexible allocation rules, that reflect the actual knowledge on the phenomenon before performing the experiment. The second main advantage of the Bayesian approach is that it addresses additional unknown quantities that are not of direct scientific interest (i.e. nuisance parameters), such as the parameters that measure the variability of the data.

1.1.1 Motivating examples

Let us recall here two examples proposed in De Santis & Perone Pacifico (2004) in order to motivate the main ideas pointed out in the above section.

Example I Let us suppose that the purpose of the study is evaluating the relative effectiveness of an innovative therapy, with respect to the standard one. This problem can be formalized, for instance, as interval estimation of the difference in means of independent normal random variables with equal unknown variances, σ^2. Hence, given a confidence level $1-\alpha$, the interval based on two independent samples of sizes n_1 and n_2 has width $2t_{n-2;1-\alpha/2}S\sqrt{n_1^{-1}+n_2^{-1}}$, where S is the pooled standard deviation, $n = n_1 + n_2$ and $t_{n-2;1-\alpha/2}$ is the $1-\alpha/2$ quantile of the Student distribution with $n-2$ degrees of freedom. Note that the above quantity depends on the random variable S. To determine n_1 and n_2, the standard procedure is to require the expected width of the random interval to be less than a chosen threshold, l (see for example Beal (1989)). However, since the expected value of S depends on the unknown value of σ, a guess value of this nuisance parameter must be chosen to select values for n_1 and n_2.

Example II Let us consider the experimental situation of a clinical trial for comparing the probabilities of success (or failure) of two competing treatments. These are the unknown parameters of two independent binomial distributions, denoted by θ_1 and θ_2. For instance, let us assume we want to estimate the unknown log odds ratio using the standard $1-\alpha$ confidence interval based on two independent samples whose sizes are indicated by n_1 and n_2. The most commonly used frequentist approach is to choose the minimal sample size that guarantees the confidence interval width is not greater than l (see O'Neill (1984)). Since the width depends on the unknown parameters (θ_1, θ_2), the criterion requires preliminary guesses, say $(\theta_{D1}, \theta_{D2})$. In De Santis & Perone Pacifico (2004) it is shown that the expression for the optimal sample size and the optimal proportion of cases directly depend on $(\theta_{D1}, \theta_{D2})$, as follows. Denoting with $z_{1-\alpha/2}$ the $1-\alpha/2$ percentile of the standard normal distribution, the resulting total sample size, $n = n_1 + n_2$, is

$$n = \frac{4z_{1-\alpha/2}}{l^2}\left(\frac{1}{\frac{n_2}{n}\theta_{D2}(1-\theta_{D2})} + \frac{1}{(1-\frac{n_2}{n})\theta_{D1}(1-\theta_{D1})}\right)$$

where the optimal proportion of cases

$$\frac{n_2}{n} = \left(1 + \sqrt{\frac{\theta_{D2}(1-\theta_{D2})}{\theta_{D1}(1-\theta_{D1})}}\right)^{-1}$$

is obtained by minimizing the asymptotic variance of the maximum likelihood estimator. Hence, if the observed proportions match the initial estimates $(\theta_{D1}, \theta_{D2})$, then the width of the confidence interval would be equal to l. Otherwise, inaccurate preliminary estimates could lead to excessively wide confidence intervals, which is the typical local optimality problem of standard SSD procedures.
In summary, the above examples are helpful in showing that the standard procedures for determining the sample size are only locally optimal, even in the simplest settings.

The outline of the present chapter is as follows. After a brief review of the main contributions in literature, we introduce the general framework of the predictive Bayesian approach to SSD. In Section 1.2.2 we highlight the possibility of eliciting two distinct priors, the one is used in the design phase and the other one for final inference. This is called two–priors approach. Finally in Section 1.4 and in Section 1.5 we specifically refer to the normal and the binomial model providing applications of the proposed criteria.

1.1.2 Review

The subject of this thesis is related to the general context of Bayesian experimental designs, illustrated in an exhaustive review by Chaloner & Verdinelli (1995). Another point of reference in the literature is the handbook by Spiegelhalter et al. (2004) that is a milestone for the use of Bayesian methods in clinical trials and health-care evaluation. This reference also gains special importance thanks to the official interest recently expressed by the FDA (i.e. Food and Drug Administration) towards the Bayesian approach. In fact, in the *Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials 2006*, FDA (2006), the FDA makes explicit, once and for all, the possibility of adopting *in practice* a Bayesian approach. This document provides the guidelines for a correct use of Bayesian techniques in clinical trials, by describing the fundamental aspects of the Bayesian paradigm, highlighting its potentialities and setting specific rules for practical applications.

The SSD problem has been addressed in Bayesian literature from several perspectives. First of all, the decision-theoretic approach is probably the most rigorous one and, in a sense, the most complete one, in that it allows to formally incorporate, in a loss function, advantages and disadvantages concerning the choice of a particular sample size – see, among others, Berger (1985), Piccinato (1996), Bernardo (1997), Lindley (1997), Raiffa & Schlaifer (2000). This approach, however, presents the intrinsic concern of the specification of a loss function, which is not immediate especially when the opinion of non-statistician experts should be formalized. A further layer of complexity is due to the involvement of different interested parties, for instance patients, physicians, pharmaceutical companies, regulatory committee, etc. As argued in Joseph et al. (1997), in general these parties have completely different points of view that translate in different loss functions. This makes it intriguing but difficult to adopt a decision-theoretic approach.

From a different perspective, several SSD criteria have been proposed by Spiegelhalter & Freedman (1986), Lee & Zelen (2000) for testing problems and by Pham-Gia & Turkkan (1992), Adcock (1997) for estimation problems. In particular, the work of Joseph and colleagues (Joseph et al. (1995, 1997), Joseph & Belisle (1997), M'Lan et al. (2006)) focuses on sample size calculations with regard to the estimation of posterior credible intervals (more specifically highest posterior density intervals) adopting a Bayesian approach that makes full use of the available prior information. In summary, the Authors define several SSD criteria in terms of the average coverage probability or the average length of intervals of posterior credible sets over all possible data sets, weighted by the predictive distribution. These criteria have been first proposed in Joseph et al. (1995) with applications to SSD for binomial proportions. Then in Joseph et al. (1997) the Authors address the case of the difference between two binomial proportions with particular attention to the *mixed Bayesian/likelihood methods* that uses the prior distribution to derive the predictive distribution of the data, founding the final inference on the likelihood only. Adopting Spiegelhalter et al. (2004)'s terminology, in Section 2.2 we name this approach *hybrid classical-Bayesian* and we illustrate it in the context of SSD methods based on the power function: its strength is the possibility to connect the Bayesian account for prior uncertainty in the planning step with a classical final inference. The Authors underline that in some situations this may be quite appropriate, as there may be substantial prior information that cannot be included in the final report for regulatory limitations. Furthermore in Joseph & Belisle (1997) SSD for normal means and difference between normal means is considered and finally a more recent work by M'Lan et al. (2006) deals with case control studies, extending the methodology first presented in De Santis et al. (2004). Among the most recent contributions we also cite Clarke & Yuan (2006), Sahu & Smith (2006) and some papers related to the issue of robustness with respect to prior specification, which we address in Chapter 3: among others, DasGupta & Mukhopadhyay (1994), De

Santis (2006), Brutti & De Santis (2008), Brutti et al. (2008*b*).

In this work we refer quite often to the *simulation-based approach* proposed by Wang & Gelfand (2002): in the Authors' words this approach "sacrifices explicit SSD formulas and is computationally intensive but is feasible for at least a portion of the wide range of hierarchical models which dominate the current Bayesian landscape". We actually resort to this framework whenever it is not possible to obtain analytic expressions for the quantities involved in the SSD calculations. Moreover, in principle, the simulation–based approach allows one to extend the methodology proposed in this thesis to more complex models.

1.2 Predictive Bayesian SSD

1.2.1 Preliminaries

Let us suppose we want to carry out a clinical trial to estimate a parameter of interest θ. Without loss of generality let us assume that the experiment is defined *successful* if it yields evidence that θ is larger than a given threshold δ. Note that in a Phase III trial θ represents a measure of comparison between two treatments and this setting reduces to the framework of a superiority trial. Furthermore let us assume that pre-experimental information on θ is available. For instance we may want to take into account the information provided by the results of a previous study or the opinion of some expert clinicians about the experimental treatment. As already mentioned in Section 1.1, according to a Bayesian perspective initial information can be formalized by specifying a prior probability distribution π_A for θ.

Let us consider the random sample $\mathbf{Y_n} = (Y_1, ..., Y_n)$, where $Y_i \sim f(\cdot;\theta)$ is the random variable associated to the effectiveness of the experimental treatment. Let us assume for the moment a prefixed number n of patients to be recruited. Once the trial has been performed, the observed sample $\mathbf{y_n} = (y_1, ..., y_n)$, which is a realization of $\mathbf{Y_n}$, is available. We denote the corresponding likelihood by $f(\mathbf{y_n};\theta)$. Then, according to the Bayesian paradigm, inference is based on the posterior distribution that follows from Bayes theorem:

$$\pi_A(\theta|\mathbf{y_n}) = \frac{\pi_A(\theta) f(\mathbf{y_n};\theta)}{m_A(\mathbf{y_n})} \tag{1.1}$$

where the denominator is the marginal distribution $m_A(\mathbf{y_n}) = \int_\Theta \pi_A(\theta) f(\mathbf{y_n};\theta) d\theta$ and Θ denotes the parameter space. Let us assume that we are interested in the posterior quantity of interest, defined as:

$$\rho_{\pi_A}(\theta|\mathbf{y_n}) = \int_{\Theta} g(\theta)\pi_A(\theta|\mathbf{y_n})d\theta. \tag{1.2}$$

Now, according to the choice of the function $g(\cdot)$ we get different summaries of the posterior distribution; in particular we focus on the following two alternatives:

a. if $g(\cdot)$ is the identity function, i.e. $g(\theta) = \theta$, we obtain the posterior expected value $\rho_{\pi_A}(\theta|\mathbf{y_n}) = E_{\pi_A}(\theta|\mathbf{y_n})$,

b. if $g(\cdot)$ is the indicator function of a given set H, i.e. $g(\theta) = I_H = \begin{cases} 1 & \theta \in H \\ 0 & \text{otherwise} \end{cases}$, we obtain the posterior probability $\rho_{\pi_A}(\theta|\mathbf{y_n}) = P_{\pi_A}(\theta \in H|\mathbf{y_n})$.

Since by definition an experiment is successful if it provides evidence of a large value of θ, it is reasonable to choose a set of this kind: $H = \{\theta : \theta > \delta\}$, where δ is a minimally clinical relevant threshold. Although the introduction of the function $g(\cdot)$ apparently involves a slight complication, in the following it turns out to be helpful in providing a unifying framework, that allows one to consider suitable transformations of the parameter (see Section 1.5.1 and Section 1.5.2). Moreover, this formulation is used in Chapter 3 to define robust SSD criteria (see in particular Section 3.2.4). Table 1.1 summarizes different choices of $g(\cdot)$ and the resulting posterior quantities of interest, that are considered in the present Chapter.

We finally need to remark that several SSD criteria, proposed for instance in Joseph et al. (1995) and Joseph et al. (1997), are based on posterior credible intervals, that actually do not appear in Table 1.1. In Chapter 5, following the approach proposed by Brutti & De Santis (2008), we adopt the credible interval as posterior quantity when dealing with equivalence trials. By the moment, we focus on the two options *a.* and *b.* only, as specified above.

Let us go back now to the main focus of this work. As we said in Section 1.1, planning the optimal sample size is a pre-experimental problem: hence, to determine the optimal sample size n^*, before the experiment we have to deal with the random sample $\mathbf{Y_n} = (Y_1, ..., Y_n)$. In particular the posterior quantity of interest $\rho_{\pi_A}(\theta|\mathbf{Y_n})$ is a function of the random data and, consequently, it is random as well. Thus, in order to take into account the randomness of the data using their marginal distribution, we need to introduce SSD criteria based on predictive summaries of $\rho_{\pi_A}(\theta|\mathbf{Y_n})$. Adopting a conditional approach as in the frequentist context, it is possible to prefix a *design value* θ_D, that is a guess value for the parameter representing the objective of the experiment or, in other words, the target effect to be detected. In this case the

parameter θ	**function of the parameter** $g(\theta)$	$\rho_{\pi_A}(\theta\|\mathbf{y_n})$
$\theta \in A \subseteq \mathbb{R}$	$g(\theta) = \theta$	$E_{\pi_A}(\theta\|y_n)$
$\theta \in A \subseteq \mathbb{R}$	$g(\theta) = I_H(\theta)$	$P_{\pi_A}(\theta \in H\|y_n)$
$\theta \in [0,1]$	$g(\theta) = \log(\theta/(1-\theta)) = \psi$	$E_{\pi_A}(\psi\|y_n)$
$\theta \in [0,1]$	$g(\theta) = I_H(\psi)$	$P_{\pi_A}(\psi \in H\|y_n)$
$\theta = (\theta_1, \theta_1) \in [0,1] \times [0,1]$	$g(\theta) = \log\left(\frac{\theta_1(1-\theta_2)}{(1-\theta_1)\theta_2}\right) = \varphi$	$E_{\pi_A}(\varphi\|y_n)$
$\theta = (\theta_1, \theta_1) \in [0,1] \times [0,1]$	$g(\theta) = I_H(\varphi)$	$P_{\pi_A}(\varphi \in H\|y_n)$

Table 1.1: Posterior quantities of interest according to the choice of $g(\theta)$

predictive summaries are computed with respect to the sampling density $f_n(\cdot;\theta_D)$. However, it is possible to model uncertainty on θ_D by specifying a prior probability distribution π_D for θ, that is also called *design prior*. As discussed in next section, in principle it can be distinct from the *analysis prior* π_A. The prior distribution π_D is used to average the likelihood, yielding the marginal predictive distribution

$$m_D(\mathbf{y_n}) = \int_\theta f(\mathbf{y_n};\theta)\pi_D(\theta)d\theta. \tag{1.3}$$

Notice that the sampling distribution of the data $f_n(\cdot;\theta_D)$ arises as a special case of $m_D(\cdot)$ when a point-mass design prior on the single value θ_D is chosen. In this sense $m_D(\cdot)$ generalizes $f_n(\cdot;\theta_D)$.

1.2.2 *Two-priors approach*

As pointed out in the previous section, when adopting a Bayesian approach we need to specify a prior distribution for computing both the posterior distribution and the predictive distribution. In general most of the Bayesian SSD criteria use the same prior distribution (see, among others, Lindley (1997), Raiffa & Schlaifer (2000)). However, several authors have argued that two priors should be used, due to the conceptual distinction between the two different roles the prior distribution is employed in: on the one hand the *design prior* models uncertainty on unknown parameter and it is used to obtain the predictive distribution (as in (1.3)); on the other hand the *analysis prior* models pre-experimental information and it is used to obtain the posterior distribution. In principle these two priors do not necessarily have to coincide. We therefore refer to this approach as *two-priors approach.*

The possibility of using different priors for design and estimation was first acknowledged in Tsutakawa (1972). The Author justified the apparent inconsistency of this innovative idea providing technical reasons: in his words, "using a design prior with variance much larger than believed reasonable is likely to lead to a wasteful experiment", while it is pretty common to consider non-informative priors for final inference. After this 'pioneer' paper, this concept has been refined by Etzioni & Kadane (1993). The motivating idea of this article is that the party performing the experiment and the party evaluating the experimental data do not necessarily have to be the same. Sometimes, even if they have common goals, their priors may be different. This is the sense of the title of the paper, *Optimal experimental design for another's analysis.* And this also responds to the point emphasized in Spiegelhalter & Freedman (1988): reviewers and consumers, rather than experimenters, ultimately determine whether new treatments are adopted in clinical practice; therefore inference should convince those evaluating medical trials, despite the prior opinion of those performing the trial.

In the most recent literature the use of two priors has been considered in a paper by Wang & Gelfand (2002): the Authors provide an exhaustive formulation of this approach, that has constituted the paradigm for a set of following works, among others Sahu & Smith (2006), De Santis (2006, 2007), Brutti & De Santis (2008), Sambucini (2008), Brutti et al. (2008*b*). Wang and Gelfand point out that it is convenient to choose a relatively non–informative analysis prior – that they call *'fitting'*, since it is used to fit the model once the data are obtained – because in general it is preferable to let the data drive inference. On the other hand, the design prior – *'sampling'* prior in their terminology – represents the scenario we expect to observe and in this sense it must be chosen to be informative. Moreover in this way one can play with different scenarios and compare the results: this is what the Authors mean by the expression *'what if'* spirit.

In conclusion, we find convincing the idea of the two priors and consequently in this thesis we adopt this approach. This also guarantees a substantive advantage in terms of flexibility and interpretability. Finally, as we argued in the previous section, the two–priors approach also constitutes a general framework including as special cases both the hybrid classical-Bayesian (described in Spiegelhalter et al. (2004)) and the classical approach. This concept is further discussed in Chapter 2, with regard to the proposed interpretation of the Predictive Bayesian Power function (defined in Section 2.2.3) as a generalized form of power.

1.2.3 Criteria

In this Section we recall and generalize the predictive Bayesian SSD criteria proposed in Brutti et al. (2008*b*). Given that the objective of the trial is to observe a large value of $\rho_{\pi_A}(\theta|\mathbf{y_n})$ (as we assumed in Section 1.2.1), we want to set suitable predictive criteria in order to control the posterior distribution through the random quantity $\rho_{\pi_A}(\theta|\mathbf{Y_n})$. As mentioned in Section 1.2.1, these criteria are based on summaries of the predictive distribution of (1.3). According to the choice of the summary, we define for instance:

1. **Predictive expectation criterion.** Let

$$e_n = \mathbb{E}_{m_D}[\rho_{\pi_A}(\theta|\mathbf{Y_n})] \tag{1.4}$$

be the expected value of $\rho_{\pi_A}(\theta|\mathbf{Y_n})$ with respect to m_D. Given a suitable threshold η_e, the chosen sample size is then

$$n_e^* = \min\left\{n \in \mathbb{N} : \ e_n > \eta_e\right\}. \tag{1.5}$$

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

2. **Predictive probability criterion.** Consider the predictive probability of obtaining a successful experiment:

$$p_n = \mathbb{P}_{m_D}[A_n] = \int_{A_n} m_D(y_n) dy_n, \tag{1.6}$$

where $\mathbb{P}_{m_D}$ is the predictive probability measure associated to m_D and A_n the subset of the sample space that contains all the samples yielding a successful experiment at level γ:

$$A_n = \left\{y_n : \ \rho_{\pi_A}(\theta|\mathbf{Y_n}) > \gamma\right\}.$$

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

$$n_p^* = \min\left\{n \in \mathbb{N} : \ p_n > \eta_p\right\}, \qquad \eta_p \in (0,1). \tag{1.7}$$

As we will show in Section 2.2.3 for $\rho_{\pi_A}(\theta|\mathbf{Y_n}) = P(\theta > \delta|\mathbf{Y_n})$ (case *b.* of Section 1.2.1), p_n coincides with the *Bayesian power* defined in Spiegelhalter et al. (2004).

A technical remark: Criterion 1 guarantees an average control on the predictive distribution of $\rho_{\pi_A}(\theta|\mathbf{Y_n})$, but in general a large predictive expected value does not necessarily avoid small values of the posterior quantity of interest. On the contrary, as argued in De Santis (2006), using Criterion 2 the sampling variability is accounted for, since one directly controls the probability of observing small values of $\rho_{\pi_A}(\theta|\mathbf{Y_n})$. In Section 1.4 and 1.5 we consider examples of both methods for the normal and the binomial model respectively.

As already noticed in Section 1.2.2, the two–priors approach can be interpreted as a general framework incorporating other approaches as special cases. In particular it reduces to the hybrid classical-Bayesian method when we choose a non–informative analysis prior and a proper design prior, while if the design prior is a point-mass prior centered on a design value θ_D we obtain the classical approach. This point will be further discussed in Chapter 2 (see in particular Table 2.3). Furthermore, it is interesting to point out that generally, at least in standard models, a non–informative analysis prior leads to a proper posterior. Conversely, a non–informative improper design prior cannot be employed since the corresponding marginal distribution of the data, m_D, is undetermined. See, for instance, De Santis (2007) for discussion on this point.

1.3 Choice of the thresholds η_e and η_p

Given the definition of the SSD criteria in the above section, it is straightforward to notice that the existence and the actual values of the optimal sample sizes n_e^* and n_p^* crucially depend on the interplay between the thresholds η_e and η_p, as well as on the choice of δ and of the design prior parameters. Since δ is defined as a minimally clinical relevant threshold, we assume that it is suggested for example by an expert and in this sense it is problem specific. Hence, given the scenario represented by the design parameter – and bearing in mind that several alternative scenarios can be compared – what we need is a criterion for setting the thresholds η_e and η_p involved in (1.5) and (1.7).

A reasonable option, suggested in Brutti et al. (2008*b*), relies on the following procedure. First of all we notice that, under the assumptions of Section 1.2.1, the predictive summaries e_n and p_n defined in (1.4) and (1.6) are increasing functions of n and they converge to the limiting quantities respectively denoted by e_∞ and p_∞, as the sample size n diverges. Without loss of generality, let us restrict ourselves to case b in which the posterior quantity of interest is the probability $P_{\pi_A}(\theta > \delta|y_n)$. Hence, it is quite intuitive that e_∞ and p_∞ equal 1 only when the design prior is a point-mass centered on a design value θ_D larger than δ; this implies that in general they are smaller than 1 for any finite n_D. Hence, if η_e and η_p are chosen as prefixed thresholds representing the trial objective, the optimal sample size can be chosen using (1.5) or (1.7) respectively; nevertheless it may happen that the optimization problem is not well posed, whenever $\eta_e > e_\infty$ or $\eta_p > p_\infty$. Since the actual ranges of the predictive quantities (e_1, e_∞) and (p_1, p_∞) heavily depend on the design parameters and on δ, in order to overcome this problem the following alternatives are available:

- it is possible to question the prefixed design scenario and change the design parameters;
- if there is evidence that the trial cannot meet its objectives, we can decide not to start it at all (this is more clear, for instance, in a re-estimation set-up, where we can decide to stop for futility);
- otherwise, a different scale can be considered, adopting the maximum achievable value as a point of reference and picking η_e (respectively η_p) as a prespecified percentage $\beta \in (0, 1]$ of e_∞ (respectively p_∞), in order to ensure the existence of the corresponding optimal sample sizes.

In summary, we only need to derive e_∞ and p_∞. For the sake of brevity, we now focus on p_n only, but as we show later, e_n and p_n are asymptotically equivalent.

First of all in order to formalize the problem better, let us define the following quantity:

$$\zeta_n = \zeta_n(x) = P_{\pi_A}(\theta > \delta | x)$$

where we set $x = y_n$ to simplify the notation. Then, from (1.6), we have

$$\begin{aligned} \lim_{n\to\infty} p_n &= \lim_{n\to\infty} \mathbb{P}_{m_D}\{\zeta_n > \gamma\} = \lim_{n\to\infty} \mathbb{E}_{m_D}\left\{I_{(\gamma,1]}(\zeta_n)\right\} = \\ &= \lim_{n\to\infty} \int_{\mathbb{R}} I_{(\gamma,1]}(\zeta_n(x)) m_D(x) dx = \int_{\mathbb{R}} \lim_{n\to\infty} \left[I_{(\gamma,1]}(\zeta_n(x)) m_D(x)\right] dx \qquad (1.8) \end{aligned}$$

where the last equality comes from an application of the dominated convergence theorem. Now, since for any regular π_A the posterior distribution is asymptotically concentrated on x, we obtain

$$\lim_{n\to\infty} \zeta_n = I_{(\delta,\infty)}(x) \qquad (1.9)$$

and consequently

$$\lim_{n\to\infty} I_{(\gamma,1]}(\zeta_n(x)) = I_{(\gamma,1]}\left(I_{(\delta,\infty)}(x)\right) = I_{(\delta,\infty)}(x).$$

Thus, whenever $m_D(\cdot) \to \pi_D(\cdot)$, the inner limit in equation (1.8) is equal to

$$\lim_{n\to\infty} \left[I_{(\gamma,1]}(\zeta_n(x)) m_D(x) \right] = I_{(\delta,\infty)}(x)\pi_D(x), \tag{1.10}$$

which does not depend on γ. Furthermore, combining the definition of e_n in equation (1.4) with the result in equation (1.9), we see that $\lim\limits_{n\to\infty} e_n = \lim\limits_{n\to\infty} p_n$.

In Section 3.5 we extend the just described procedure to the robust criteria introduced in Chapter 3. Similarly in Section 4.2.4 the asymptotic behaviour of the expected posterior probability is discussed, in case the analysis prior is a mixture of prior distributions derived from several sources.

1.4 Results for normal model

Let us assume that the 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 and y_n its observed value. However, the same basic model provides an approximation that can be used, for instance, for binary data – with θ denoting the log odds ratio – and for survival data – with θ denoting the log hazard ratio – (see Spiegelhalter et al. (2004), Sections 2.4.1 and 2.4.2). For computational simplicity we adopt conjugate priors. Thus we assume that π_A is a normal density of mean θ_A and variance σ^2/n_A, where n_A is the prior sample size. From standard Bayesian conjugate analysis it follows that the resulting posterior distribution is again a normal density with mean

$$E_{\pi_A}(\theta|y_n) = \frac{n_A\theta_A + ny_n}{n_A + n} \tag{1.11}$$

and variance

$$V_{\pi_A}(\theta|y_n) = \frac{\sigma^2}{n_A + n}. \tag{1.12}$$

According to the two options listed in Section 1.2.1 we have as posterior quantities of interest:

a. the posterior expectation $\rho_{\pi_A}(\theta|\mathbf{y_n}) = \frac{n_A\theta_A + ny_n}{n_A+n}$,

b. the posterior probability $\rho_{\pi_A}(\theta|\mathbf{y_n}) = 1 - \Phi\left(\frac{\delta - E_A(\theta|y_n)}{\sqrt{V_A(\theta|y_n)}}\right)$, where Φ denotes the cumulative distribution function of a standard normal.

Furthermore we assume that the design prior is $\pi_D(\theta) = N(\theta|\theta_D, \sigma^2/n_D)$. According to (1.3), the marginal predictive distribution induced by π_D is a normal density of mean θ_D and variance $\sigma^2(1/n + 1/n_D)$. Hence, given the above results and using (1.4) and (1.6), it is straightforward to derive e_n and p_n, respectively for the two choices of $\rho_{\pi_A}(\theta|\mathbf{y_n})$:

1. *a.*

$$\mathbb{E}_{m_D}\left[\frac{n_A\theta_A + ny_n}{n_A + n}\right] = \frac{n_A\theta_A + n\theta_D}{n_A + n},$$

for the linearity of the expected value;

b.

$$\mathbb{E}_{m_D}\left[1 - \Phi\left(\frac{\delta - E_A(\theta|y_n)}{\sqrt{V_A(\theta|y_n)}}\right)\right],$$

where the expected value cannot be derived analytically, but can be easily computed by simulation;

2. *a.*

$$\mathbb{P}_{m_D}\left[\frac{n_A\theta_A + ny_n}{n_A + n} > \gamma\right] = 1 - \Phi\left(\frac{\frac{\gamma(n_A+n)-n_A\theta_A}{n} - \theta_D}{\sigma\sqrt{n^{-1} + n_D^{-1}}}\right),$$

b.

$$\mathbb{P}_{m_D}\left[1 - \Phi\left(\frac{\delta - E_A(\theta|y_n)}{\sqrt{V_A(\theta|y_n)}}\right) > \gamma\right] =$$

$$= \quad 1 - \Phi\left(\frac{n^{-1}\left\{(n_A + n)(\delta - \sigma(n_A + n)^{-1/2}z_{1-\gamma}) - \theta_A n_A\right\} - \theta_D}{\sigma\sqrt{n^{-1} + n_D^{-1}}}\right),$$

where $z_{1-\gamma}$ denotes the quantile of a standard normal at level $1 - \gamma$.

The corresponding four SSD predictive criteria immediately result, according to (1.5) and (1.7). In practice, the optimal sample size is determined by computing one of the above quantities for increasing values of n and by picking the minimum sample size that guarantees to reach a prefixed threshold. This procedure can be effectively represented by a plot of the chosen predictive summary with respect to n, as we illustrate in the application of the following Section.

Example 1: Bayesian SSD for the normal model (CANCER) First of all let us introduce an example proposed in Spiegelhalter et al. (2004) (see Examples 2.6 and 6.2 for details). A randomized controlled trial is designed for testing the effects difference of two competing cancer treatments, in terms of mortality. Hence, the log hazard ratio of death is chosen as a measure to compare the events occurring in two randomized arms and a normal approximation is used. The trial was design to have a 80% power to detect a log hazard ratio $\theta_D = 0.56$, equivalent to a raise of 5-year survival from 20 to 40 per cent in favour of the new treatment. The Authors considered a design prior centered on the guessed value θ_D and with 0.05 probability that θ is less than zero, indicating that the old treatment is better than the new one. This results in a design prior sample size $n_D = 34.5$ and, overall, in a design density that represents optimism towards the new treatment. The prior is then employed to average the classical power curve and to obtain a hybrid classical Bayes power to be compared with the standard procedure, which is equivalent to using a non–informative analysis prior for θ. We here extend Spiegelhalter et al. (2004)'s example first of all by adopting the two–priors approach described in Section 1.2.2. We introduce an analysis prior π_A that is a normal density centered on $\theta_A = 0$, expressing equivalence between old and new treatments, and variance such that the probability that θ is greater than θ_D is equal to a chosen value α. This choice yields a prior sample size n_A equal to $(2z_{1-\alpha}/\theta_D)^2$. Note that, the smaller the values of α and of $|\theta_D|$, the more sceptical the analysis prior results. Of course an equivalent way to define a sceptical base prior is to fix θ_D and then set θ_A 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 still allow for a relatively high

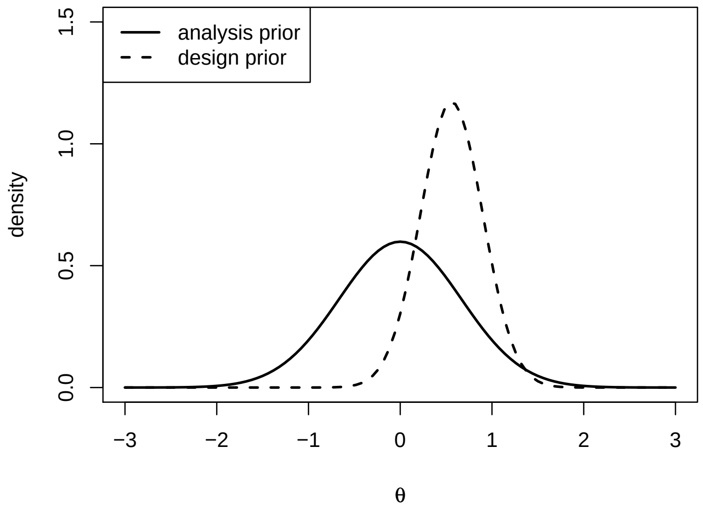


Figure 1.1: $\pi_A(\theta) = N(\theta|\theta_A = 0, \sigma^2/n_A = 4/9)$ (solid line); $\pi_D(\theta) = N(\theta|\theta_D = 0.56, \sigma^2/n_D = 4/34.5)$ (dashed line)

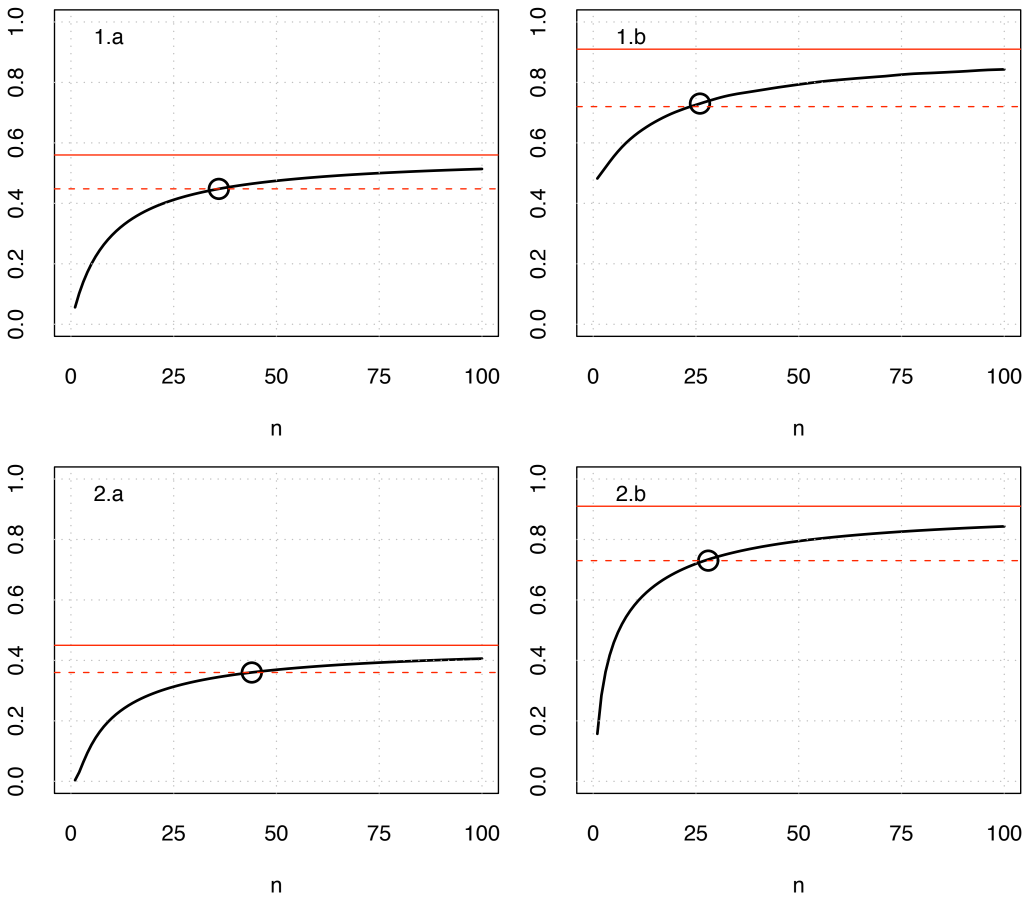


Figure 1.2: SSD criteria using the priors of Figure 1.1, with $\delta = 0.1$, $\gamma = 0.6$.

uncertainty, corresponding to a low value of the prior sample size, namely $n_A = 9$. The analysis prior is therefore less informative than the design prior, as shown in Figure 1.1. In addition, we assume a minimally clinical significant difference $\delta = 0.1$, corresponding to a raise in the survival rate from 20 to 23.3 per cent. This example is developed in next chapter with specific reference to the power-based SSD methods. A further extension is then proposed in Chapter 3, where we apply the robust SSD criteria in the same setting. In Figure 1.2 we represent the four alternative predictive quantities with respect to the sample size n. The horizontal continuous line indicates the maximum reachable value for each considered predictive summary, given δ and the design parameters. As proposed in Section 1.3, the thresholds η_e and η_p are chosen at a prefixed percentage of e_∞ and p_∞; in this case we set $\beta = 80\%$ (see Table 1.2). Finally η_e and η_p are represented by the horizontal dashed lines and the optimal sample sizes are circled in correspondence of these thresholds (and bolded in Table 1.4) .

θ_D	e_∞ (η_e)		p_∞ (η_p)	
	1.a	*1.b*	*2.a*	*2.b*
0.30	0.30 (0.24)	0.72 (0.58)	0.19 (0.15)	0.72 (0.58)
0.56	0.56 (0.45)	0.91 (0.73)	0.45 (0.36)	0.91 (0.73)
0.80	0.80 (0.64)	0.98 (0.78)	0.72 (0.58)	0.98 (0.78)

Table 1.2: e_∞, p_∞ and in brackets the corresponding thresholds η_e and η_p for each design prior mean, given $\delta = 0.1$ and $\gamma = 0.6$,

		1.a				**1.b**			
		n_A				n_A			
	θ_A	5	9	15	30	5	9	15	30
$\theta_D = 0.3$	-0.2	38	67	111	221	30	35	53	87
	-0.1	30	52	86	171	26	28	42	65
	0	22	38	61	121	20	21	31	44
	0.1	15	23	36	71	14	14	18	20
	0.2	4	8	12	21	4	3	1	1
$\theta_D = 0.56$	-0.2	31	53	88	174	29	35	44	65
	-0.1	27	45	74	148	26	31	38	53
	0	22	**38**	61	121	24	**27**	31	41
	0.1	18	30	48	94	22	23	25	29
	0.2	15	22	35	67	19	19	18	17
	0.3	9	15	22	41	17	15	11	3
	0.4	1	4	8	14	14	9	1	1
$\theta_D = 0.8$	-0.2	28	49	80	158	26	31	39	65
	-0.1	25	43	70	140	24	28	35	56
	0	22	38	61	121	22	25	30	44
	0.1	20	32	52	102	20	22	25	29
	0.2	17	27	43	83	18	18	18	5
	0.3	14	21	33	65	15	13	9	1
	0.4	11	16	24	46	12	8	1	1

Table 1.3: Optimal sample sizes according to the Predictive Expectation Criterion, for several choices of the design and analysis prior parameters, given $\delta = 0.1$ and $n_D = 34.5$ and the corresponding thresholds η_e given in Table 1.2

		2.a				**2.b**			
		n_A				n_A			
	θ_A	5	9	15	30	5	9	15	30
$\theta_D = 0.3$	-0.2	14	42	126	356	51	60	81	125
	-0.1	4	30	99	298	46	51	68	98
	0	1	21	74	240	41	42	54	71
	0.1	1	15	51	182	36	32	40	45
	0.2	1	10	35	126	30	22	23	1
$\theta_D = 0.56$	-0.2	31	68	127	277	31	38	47	70
	-0.1	26	57	109	239	28	33	40	56
	0	21	**47**	90	202	26	**29**	33	42
	0.1	17	37	72	164	23	24	26	28
	0.2	13	27	54	127	20	19	18	12
	0.3	9	18	37	90	17	14	1	20
	0.4	1	12	21	54	14	1	17	61
$\theta_D = 0.8$	-0.2	44	66	101	190	27	32	44	70
	-0.1	41	59	90	167	25	29	39	61
	0	38	53	79	145	23	26	33	49
	0.1	34	47	68	123	20	21	25	27
	0.2	31	41	57	101	16	16	9	1
	0.3	26	36	47	79	12	6	1	1
	0.4	21	29	38	57	6	6	1	1

Table 1.4: Optimal sample sizes according to the Predictive Probability Criterion, for several choices of the design and analysis prior parameters, given $\delta = 0.1$, $\gamma = 0.6$ and $n_D = 34.5$ and the corresponding thresholds η_p given in Table 1.2

In Table 1.4 we report the optimal sample sizes obtained for several combinations of the design prior mean and of the analysis prior parameters. It is evident that for each fixed value of θ_D, given a certain prior sample size n_A, the more sceptical the analysis prior, the larger the number of units required. At the same time, when we choose a more optimistic design value, the corresponding predictive distribution of the data is enthusiastic as well and a (uniformly) smaller number of observations is sufficient to achieve the study objective. Note that for each block of the table we have a different threshold η_e or η_p, depending (through e_∞ and p_∞) on the design parameters and (eventually) on δ (see Table 1.2).

1.5 Results for the binomial model

In the present section we assume that the parameter of interest θ is the probability of success of a given treatment. As in Section 1.2.1, the experiment is assumed to be successful if it provides evidence that θ is sufficiently large. Hence, we are in the setting of a Phase II trial and, more precisely, of an efficacy study. Alternatively we could consider for instance the probability of failure of a treatment in a safety study: in this case the trial would be aimed at bringing evidence of a sufficiently small value of θ.

Let us consider a random sample $\mathbf{Y_n} = (Y_1, ..., Y_n)$, where Y_i is a binary random variable associated to the success of the experimental treatment for the i-th patient, i.e. $Y_i \sim Bernoulli(\theta)$. Of course the definition of success is problem specific. Once the experiment is performed the statistic we are interested in is the total number of successes $s_n = \sum_{i=1}^{n} y_i$. We denote the likelihood by $f(s_n; \theta)$. Note that the corresponding random variable S_n, given the unknown parameter θ, is a binomial random variable with parameters (n, θ). In the following section we distinguish the standard case in which we directly focus on the probability of success θ, from the slightly different case in which a suitable transformation on the log odds scale is considered. Finally in Section 1.5.2, we cope with case control studies and the log odds ratio is employed as a measure of comparison between two competing treatments.

1.5.1 One sample

Let us start considering the standard setting in which we observe one sample of patients and we focus on the probability of success θ as a parameter of interest. As in Section 1.4, for the sake of simplicity we adopt the conjugate prior for the binomial model. Hence, we have that $\pi_A(\theta) = Beta(\theta|\alpha, \beta)$ where $Beta(\cdot|\alpha, \beta)$ denotes a beta density of parameters (α, β). This choice is motivated by (i) analytical tractability, (ii) shape flexibility of the beta distribution. that allows the experimenter to represent and formalize very different prior beliefs in a relatively straightforward way (see Spiegelhalter et al. (2004) for discussion and examples). From standard results of conjugate analysis (see, for instance, Bernardo & Smith (1994)) the posterior density of θ is still a beta density with updated parameters, namely

$$\pi_A(\theta|s_n) = Beta(\theta|\alpha + s_n, \beta + n - s_n). \tag{1.13}$$

Based on the posterior distribution defined above, we can explicitly derive the following posterior quantities, according to the definitions of Section 1.2.1:

a. the posterior expectation $\rho_{\pi_A}(\theta|\mathbf{y_n}) = \frac{\alpha+s_n}{\alpha+\beta+n}$,

b. the posterior probability $\rho_{\pi_A}(\theta|\mathbf{y_n}) = 1 - F_{B(\alpha+s_n,\beta+n-s_n)}(\delta)$, where $F_{B(\alpha,\beta)}$ denotes the cumulative distribution function of a beta density of parameters (α,β).

At this point we choose as design prior a beta density of parameters (α_D,β_D), to be specified according to the goal of the trial. Consequently, it is well known that the resulting marginal m_D is a betabinomial distribution of parameters (α_D,β_D,n). Again, the marginal distribution is used for computing the predictive summaries. We adopt here the same scheme as in Section 1.4 and list the following four options according to the choice of the posterior quantity and of the predictive summary:

1. *a.*

$$\mathbb{E}_{m_D}\left[\frac{\alpha_A+s_n}{\alpha_A+\beta_A+n}\right] = \frac{\alpha_A+n\frac{\alpha_D}{\alpha_D+\beta_D}}{\alpha_A+\beta_A+n},$$

where we used the linear property of the expected value and the expression of the mean of a betabinomial distribution;

b.

$$\mathbb{E}_{m_D}\left[1-F_{B(\alpha+s_n,\beta+n-s_n)}(\delta)\right] = \sum_{k=0}^{n}\left[1-F_{B(\alpha+k,\beta+n-k)}(\delta)\right]p_{m_D}(k),$$

where $p_{m_D}(k) = \left(\begin{array}{c} n \\ k \end{array}\right)\frac{B(\alpha_D+k,\beta_D+n-k)}{B(\alpha_D,\beta_D)}$ is the betabinomial probability of k successes out of n patients and $B(\alpha,\beta)$ denotes the beta function;

2. *a.*

$$\mathbb{P}_{m_D}\left[\frac{\alpha_A+s_n}{\alpha_A+\beta_A+n} > \gamma\right] = \sum_{k=\lceil\gamma(\alpha_A+\beta_A+n)-\alpha_A\rceil}^{n} p_{m_D}(k),$$

b.

$$\mathbb{P}_{m_D}\left[1-F_{B(\alpha+s_n,\beta+n-s_n)}(\delta) > \gamma\right] = \sum_{\{k:\,1-F_{B(\alpha+k,\beta+n-k)}(\delta)>\gamma\}} p_{m_D}(k),$$

Notice that in this case, due to the discrete nature of the marginal betabinomial distribution, predictive summaries reduce to summations over the number of successes, in such a way that opportune conditions are satisfied. This enables one to compute the above quantities exactly, without resorting to simulation. Again, it is straightforward to define the SSD criteria based on the above predictive summaries using (1.5) and (1.7). An application of these criteria is illustrated in the following paragraph.

Example 2: Bayesian SSD for the binomial model (DRUG) Let us consider for example an efficacy trial aimed at assessing the true response rate of a drug, illustrated in Spiegelhalter et al. (2004). Let us suppose that previous experience with similar compounds has suggested that response rates between 0.2 and 0.6 could be feasible, with an expectation around 0.4. The Authors suggest to specify a beta prior, where the parameters α and β are derived given the mean $m = 0.4$ and the standard deviation $s = 0.1$. In this way they elicit the analysis prior $\pi_A(\theta) = Beta(\theta|\alpha_A = 9.2, \beta_A = 13.8)$. Moreover, we consider different scenarios for the design prior. First of all, in order to have more informative prior distributions, we set the standard deviation equal to 0.05, and we consider, for instance, design prior means respectively equal to 0.6, 0.7, 0.8 and 0.9. Note that the higher the design mean the more enthusiastic the prior is with respect to the goal of the trial. The analysis prior and the different choices for the design prior are represented in Figure 1.3.

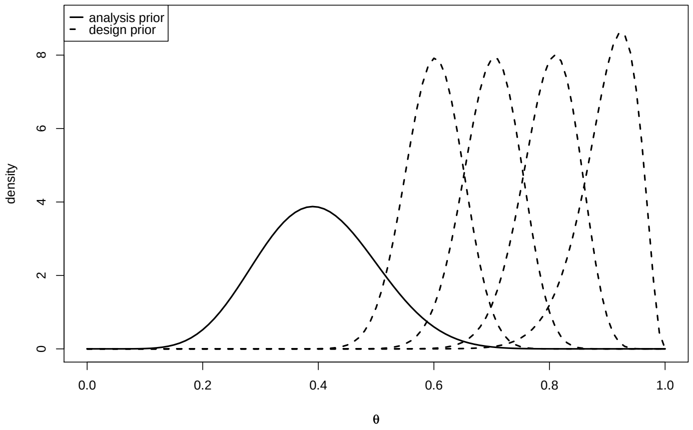


Figure 1.3: Analysis prior $\pi_A(\theta) = Beta(\theta|\alpha_A = 9.2, \beta_A = 13.8)$ and different scenarios for the design priors π_D: we choose the beta parameters $(57, 38)$, $(58.1, 24.9)$, $(50.4, 12.6)$, $(31.5, 3.5)$ to get the corresponding prior means 0.6, 0.7, 0.8, 0.9 and a standard deviation equal to 0.05

In this framework, we want to determine the optimal sample size for an efficacy trial on the same drug. For example, if we consider the predictive expectation criterion and we focus on the posterior expectation of θ, we get the results of Figure 1.4 that highlights how the behaviour of e_n changes according to the design prior. The choice of the threshold η_e is adapted to the maximum achievable value of e_n, as suggested in Section 1.3. In other words, the larger the design mean the higher e_∞, but the higher the threshold η_e as well. For instance when considering the most enthusiastic design mean ($\theta_D = 0.9$) we are also imposing a more demanding threshold ($\eta_e = 0.72$) for the selection of the optimal sample size, which results in this case $n_e^* = 42$.

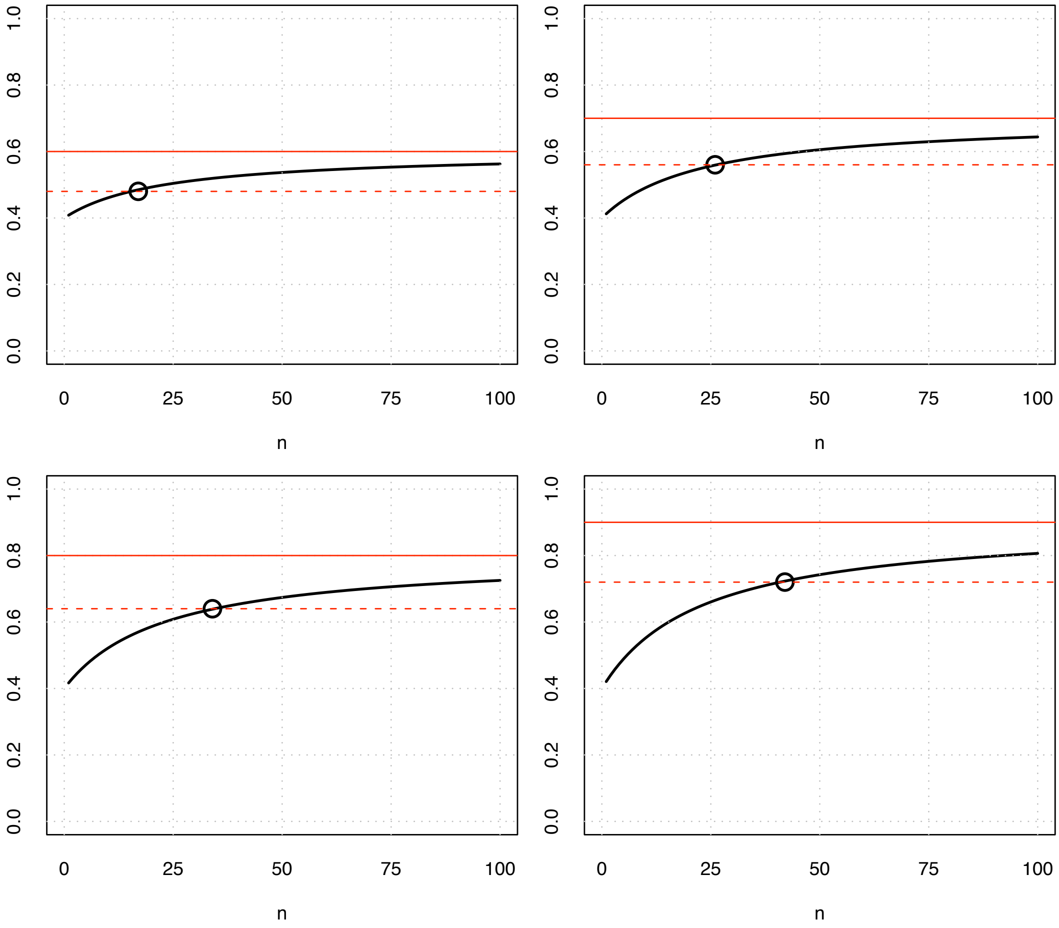


Figure 1.4: Optimal sample sizes for the different design priors of Figure 1.3, when the predictive expectation of the posterior expectation is considered with respect to n. The thresholds η_e, represented by the horizontal dashed lines, are respectively 0.48, 0.56, 0.64, 0.72 and the resulting optimal sample sizes 17, 26, 34, 42.

In Figure 1.5 option *2.b* is represented, for $\delta = 0.5$ and $\gamma = 0.8$. The plot shows the typical "sawtooth" behaviour of the predictive summary p_n with respect to n, due to the discrete nature of the betabinomial marginal distribution. In this case, given η_p, the optimal sample size n_p^* could be chosen as the minimum n that guarantees to have $p_n > \eta_p$. However, as shown in the picture, this choice could result in the paradox of selecting a sample size that satisfies a criterion that is not satisfied anymore for some greater sample size values. Hence, as suggested for instance in Sambucini (2008), we adopt a more conservative criterion that requires to select the smallest sample size n_p^* such that condition (1.7) is satisfied $\forall n > n_p^*$.

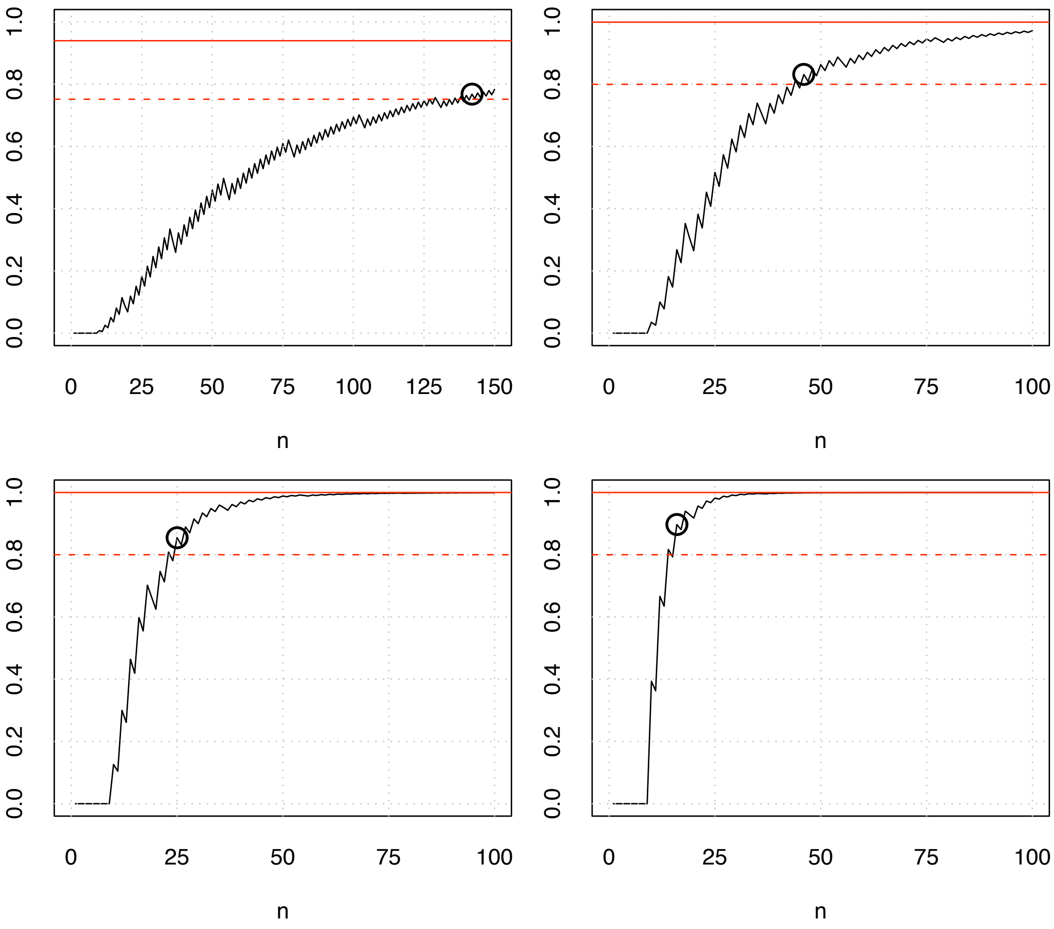


Figure 1.5: Optimal sample sizes for the different design priors of Figure 1.3, when the predictive probability of the posterior probability is considered with respect to n, with $\delta = 0.5$ and $\gamma = 0.8$. The thresholds η_p are respectively 0.75, 0.80, 0.80, 0.80 and the resulting optimal sample sizes 142, 46, 25, 16.

Log odds scale

In some circumstances, one may prefer to consider a different scale for the parameter of interest. For example, it is quite common to transform the probability of success θ on the log odds scale. Hence, we define $\psi = g(\theta) = \log(\theta/(1-\theta))$ as parameter of interest (see also Table 1.1). Note that this transformation allows one to work on the real axis, since $\psi \in \mathbb{R}$; in particular it is also possible to adopt a normal approximation for ψ, as derived in Spiegelhalter et al. (2004). In this case the procedure described in Section 1.4 applies. Nevertheless here we are interested in the framework just introduced in Section 1.5 with a binomial model and beta priors for θ. Given the posterior for θ in (1.13), it is quite straightforward to derive the posterior density of the log odds (see for example Wasserman (2004)):

$$f(\psi|y_n) = \frac{\Gamma(n+\alpha_A+\beta_A)}{\Gamma(s_n+\alpha_A)\Gamma(n-s_n+\beta_A)} \left(\frac{e^{\psi}}{1+e^{\psi}}\right)^{s_n+\alpha_A-1} \left(\frac{1}{1+e^{\psi}}\right)^{n-s_n+\beta_A+1} e^{\psi}. \tag{1.14}$$

From this exact distribution we should compute the posterior quantities of interest as defined in Section 1.2.1, although it is not possible to obtain closed-form expressions. A practical alternative, also suggested in Wasserman (2004), is the approximation of (1.14) by simulation. It is sufficient to proceed according to the following steps: (i) draw a sample from the posterior distribution of θ, (ii) apply the log odds transformation to each sampled value and finally (iii) get Monte Carlo estimates of the expectation and of the desired tail probabilities. This procedure has been implemented to compute (1.4) and (1.6): an example follows in next paragraph.

Example 2 *(continued)*: Bayesian SSD for the binomial model (DRUG) Let us suppose now that we are in the same setting described in Example 2 (page 25), but we focus on the log odds as parameter of interest. First of all we specify the analysis and the design prior for θ, for instance $\pi_A(\theta) = Beta(\theta|9.2, 13.8)$ and $\pi_D(\theta) = Beta(\theta|31.5, 3.5)$. In the left panel of Figure 1.6 we plot for example the predictive expectation of the posterior expectation of the log odds with respect to n: for a given threshold $\eta_e = 1$ we obtain $n_e^* = 44$. Given a minimally clinical relevant threshold on the log odds scale, say $\delta = 1$, we consider then the posterior probability that ψ is larger than 1. Note that this is equivalent to consider the posterior probability that θ is larger than a threshold $\delta' = exp(\delta)/(1+exp(\delta))$. The predictive probability that this quantity exceeds a given value γ (equals to 0.8, in the example) is represented in the right panel of Figure 1.6 for increasing values of the sample size. For a prefixed threshold $\eta_p = 0.8$, the optimal sample size is $n_p^* = 95$.

1.5.2 Two samples

In this Section we deal with Bayesian SSD for case control studies. First of all we need to briefly describe the general framework of a Phase III trial, whose purpose is the comparison of two competing treatments in terms of efficacy.

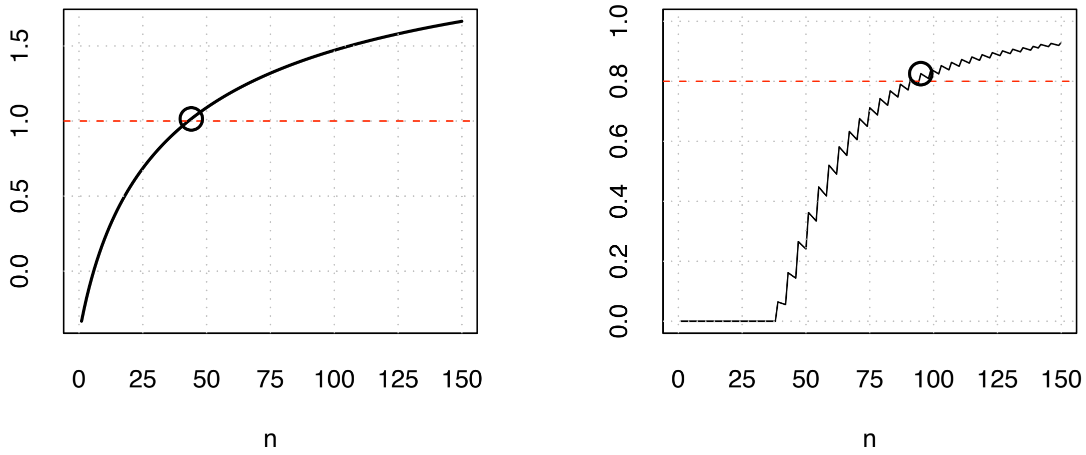


Figure 1.6: For the log odds as parameter of interest: *(left panel)* e_n with respect to n: for $\eta_e = 1$, $n_e^* = 44$. *(right panel)* p_n with respect to n, with $\delta = 1$ and $\gamma = 0.8$: for $\eta_p = 0.8$, $n_p^* = 95$.

Case control studies

A case control study is typically a controlled trial in which patients are randomly allocated in two treatment arms. The control arm (in the following indicated by 1) is treated with the standard drug, while the case arm (denoted by 2) receives the new therapy. Hence, we have respectively n_1 and n_2 patients, with $n_1 + n_2 = n$, the total number of patients to be chosen. Notice therefore that besides determining the optimal sample size a second problem is in order: in fact it is necessary to assign the units in two groups according to a reasonable criterion. One possibility, first proposed in De Santis et al. (2004), is illustrated at the end of this Section.

Let us consider $\theta = (\theta_1, \theta_2)$, with θ_i indicating the probability of an event occurring in group i, for $i = 1, 2$. Without loss of generality we consider the probability of a negative event, that is a failure such as death or disease recurrence, instead of the probability of success. Then, we choose the log odds ratio (logOR) as a measure of comparison between the two treatments effect (see again Table 1.1 for the choice of the transformation $g(\cdot)$). By definition, the logOR is

$$\varphi = \log\left(\frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_2}{1-\theta_2}}\right) = \log\left(\frac{\theta_1}{1-\theta_1}\right) - \log\left(\frac{\theta_2}{1-\theta_2}\right). \tag{1.15}$$

Note that $\theta_1 > \theta_2$ implies $\varphi > 0$, indicating that the probability of a failure under the standard treatment is larger than the one under the new experimental treatment, in other words, the new treatment is *better* than the standard one. This is then the typical setting of a superiority trial. In general, using the logOR is a standard way of reporting changes in the chances of events due to an intervention, on a scale between $-\infty$ and ∞. Note that, as discussed in Spiegelhalter et al. (2004), this is helpful to derive a normal approximation. For an application of this result with regard to the SSD problem, see Brutti et al. (2008*b*) and Example 1 in Section 1.4.

Let us report the collected data in the following 2×2 contingency table, where s_i is total number of events occurring in arm $i = 1, 2$:

	Treatment		
	Standard	**Experimental**	**Total**
Success	s_1	s_2	s_n
Failure	$n_1 - s_1$	$n_2 - s_2$	$n - s_n$
Total	n_1	n_2	n

Table 1.5: Events in a case control study.

We denote the corresponding random variable by S_i. Given the unknown parameter θ_i, S_i is a binomial random variable with parameters (n_i, θ_i). Furthermore, assuming independence of the two samples, the joint sampling distribution of (S_1, S_2) is the product of two binomial distributions.

Bayesian SSD criteria for case control studies

Let us go over the necessary steps to derive the SSD predictive criteria. For the sake of simplicity, we adopt again conjugate beta priors both for θ_1 and θ_2, as pointed out in Section 1.5. Moreover, we assume prior independence, i.e.

$$\pi_A(\theta_1, \theta_2) = Beta(\theta_1|\alpha_1, \beta_1) \cdot Beta(\theta_2|\alpha_2, \beta_2),$$

that results in posterior independence, i.e.

$$\pi_A(\theta_1, \theta_2|s_1, s_2) = Beta(\theta_1|\alpha_1 + s_1, \beta_1 + n_1 - s_1) \cdot Beta(\theta_2|\alpha_2 + s_2, \beta_2 + n_2 - s_2).$$

The posterior density of the logOR given the data has been derived in Nurminen & Mutanen (1987) and Marshall (1988):

$$\pi_A(\varphi|S_1 = s_1, S_2 = s_2) = \frac{\exp(a_2\,\varphi)}{B(a_1, b_1)B(a_2, b_2)} \int_0^1 \frac{x^{a_1+a_2-1}(1-x)^{b_1+b_2-1}}{[1 + (\exp(\varphi) - 1)x]^{a_2+b_2}} dx \quad (1.16)$$

where $a_i = \alpha_i + s_i$ and $b_i = \beta_i + n_i - s_i$ are the posterior parameters of θ_i, for $i = 1, 2$. However, whenever it is necessary to compute posterior quantities of interest, it is also possible to resort to Monte Carlo simulation (see again Wasserman (2004)).

Based on the posterior distribution given in (1.16), we define, as usual, the posterior quantities of interest such as, for instance

a. the posterior expected value: $\rho_{\pi_A}(\theta|\mathbf{y_n}) = \rho_{\pi_A}(\varphi|s_1, s_2) = E_{\pi_A}(\varphi|s_1, s_2)$,

b. the posterior probability of a given subset H: $\rho_{\pi_A}(\theta|\mathbf{y_n}) = \rho_{\pi_A}(\varphi|s_1, s_2) = P_{\pi_A}(\varphi \in H|s_1, s_2)$.

Recalling definition (1.15), we know that $\varphi > 0$ favors the innovative therapy. Hence, it is reasonable to set $H = \{\varphi : \varphi > \delta\}$, where δ is a minimally clinical relevant threshold on the logOR scale.

Again, a trial is defined *successful* if the value of the posterior quantity $\rho_{\pi_A}(\varphi|s_1, s_2)$ results larger than a reference value. However, before the experiment is performed, $\rho_{\pi_A}(\varphi|S_1, S_2)$ is a random quantity. Hence, we consider the predictive expectation or the predictive probability of $\rho_{\pi_A}(\varphi|S_1, S_2)$, as discussed in Section 1.2.3. Thus, choosing as design priors for θ_1 and θ_2 independent conjugate beta priors of parameters respectively $(\alpha_{D1}, \beta_{D1})$ and $(\alpha_{D2}, \beta_{D2})$, the following marginal distribution results

$$m_D(s_1, s_2) = m_{D_1}(s_1) \cdot m_{D_2}(s_2), \tag{1.17}$$

where $m_{D_i}(s_i)$ is a betabinomial distribution of parameters $(\alpha_{Di}, \beta_{Di}, n_i)$, for $i = 1, 2$. It is then straightforward to define the usual predictive summaries using $\rho_{\pi_A}(\varphi|S_1, S_2)$ in Equations (1.4) and (1.6). Then the SSD criteria are well defined, once the thresholds η_e and η_p are conveniently fixed (see again Section 1.3). Before illustrating an application, in the following paragraph we cope with the problem of units allocation in two randomized arms.

Allocation of observations

As anticipated at the beginning of this Section, patients allocation is naturally connected to SSD: besides selecting the optimal total number of patients we also need to decide in which proportion they should be randomly assigned either to the new treatment or to the standard one. This twofold decision obviously influences the SSD criteria definition. Technically speaking, from (1.17) it is evident that the marginal distribution – and consequently the SSD predictive criteria – depends on n_1 and n_2. Hence, in principle, for each candidate sample size n we should consider each couple (n_1, n_2) summing to n and select the one optimizing the predictive criterion. This procedure is computationally intensive and results to be impractical. As an alternative we adopt the solution, proposed in De Santis et al. (2004), of choosing (n_1, n_2) first, in such a way that the expectation of the posterior variance of the unknown parameters is equal. This rule typically yields a larger number of cases (i.e. $n_2 > n_1$), which is consistent since pre-experimental information about the control group is usually more accurate, being supported by the results of previous trials. Hence, the first step in the SSD procedure is allocation of units in two arms for each fixed n. After that, one can finally determine the minimum sample size satisfying the desired predictive criterion.

Example 3: Bayesian SSD for the log odds ratio (GREAT) We present here an example of the proposed methodology, with reference to an application described in Pocock and Spiegelhalter (1992) and discussed further in Spiegelhalter et al. (2004). Let us consider the randomized controlled trial named GREAT on a new thrombolytic therapy after myocardial infarction. The purpose of the study is to compare two competing treatments: a new drug (anistreplase) against a placebo. The outcome is thirty-day mortality rate under each treatment. We consider here the logOR scale and, according to the notation introduced in Section 1.5.2, we have that $\varphi > 0$ (that is, equivalently, OR> 1) supports the new treatment. The observed data are reported in Table 1.6; in Spiegelhalter et al. (2004) the analysis is carried out according to a Bayesian approach and it is based on the normal approximation of the log odds ratio. Anyways, since we are interested in the pre-experimental aspects of the problem, we imagine here to plan a new experiment and we use the data to elicit the prior distributions. Let us suppose we want to show a treatment difference similar to the one provided by the results actually observed in the GREAT trial: then, based on these data, we can specify the design priors.

Following Spiegelhalter et al. (2004), it is well known that the parameters a and b of a beta prior can be given a straightforward interpretation: a represents the number of events occurred in an hypothetical previous trial of size $a + b$. Consequently, b can be thought as the number of elements who did not experience any event. In the light of this meaning of the parameters, we "translate" the data in Table 1.6 into

	control treatment	**new treatment**	**tot**
death	23	13	36
no death	125	150	275
tot	163	148	311

Table 1.6: Results of GREAT study

the design prior parameters. Thus we have respectively for θ_1 and θ_2:

$$\pi_D(\theta_1) = Beta(23, 125) \quad \text{and} \quad \pi_D(\theta_2) = Beta(13, 150). \tag{1.18}$$

Furthermore we assume that the analysis priors for θ_1 and θ_2 both represent a certain degree of scepticism towards the treatment (prior mean equal to 0.167) and coincide, indicating no treatment difference:

$$\pi_A(\theta_1) = \pi_A(\theta_2) = Beta(2, 10). \tag{1.19}$$

This choice of the parameters for the Beta prior yields a standard deviation equal to 0.1. This guarantees that the analysis priors are less informative than the design ones, which is coherent with their intrinsic meaning (see Wang & Gelfand (2002) and Section 1.2.2). In Figure (1.7) we represent the analysis and design priors for θ_1 and θ_2 in the left panel and the corresponding analysis and design prior distribution of the log odds ratio φ in the left one, where the exact distribution is compared with the Monte Carlo simulated one and with the normal approximation.

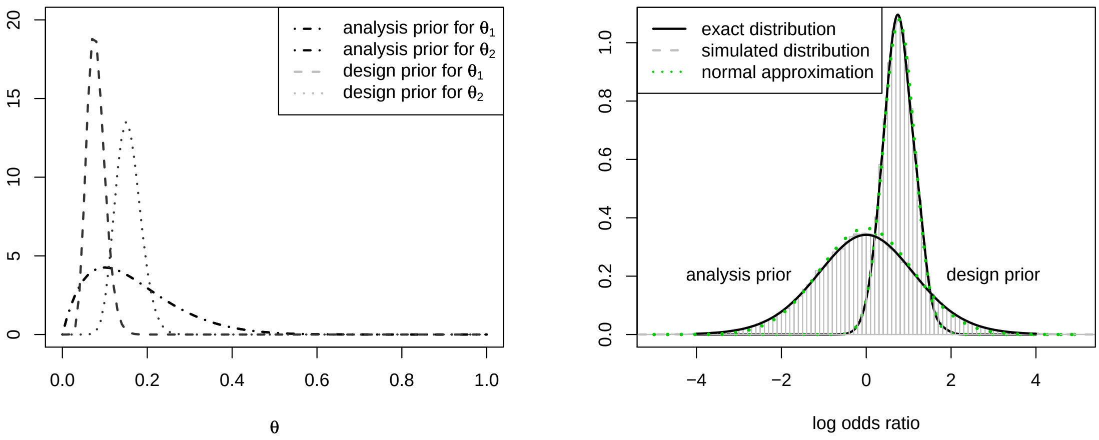


Figure 1.7: Analysis priors (continuous line) and design priors (dotted line) for θ_1 and θ_2 (*left panel*) and for φ (*right panel*).

If for instance we focus on the posterior probability that $\varphi > \delta$ and we consider the predictive expectation as n increases, we obtain the situation represented in Figure 1.8 which can be given an analogous interpretation to the one of the plots of the previous sections: the resulting optimal sample size is $n_e^* = 130$, for $\delta = 0$ and $\eta_e = 0.8$ (left panel); choosing a larger value of δ, such as 0.5, the expected probability is appreciably lower, $n_e^* = 197$ for $\eta_e = 0.6$.

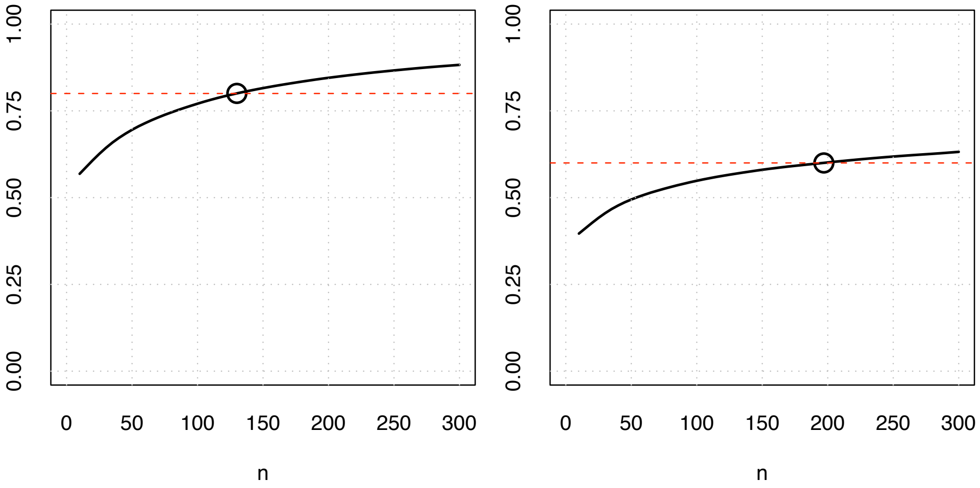


Figure 1.8: SSD using criterion 1.b, with $\delta = 0$ (left panel) and $\delta = 0.5$ (right panel).

1.6 Extensions and further developments

In the present chapter we have introduced the general framework of a predictive Bayesian approach to SSD, providing practical examples for the normal model and for the binomial model. As already noticed, a special case of the proposed methodology (corresponding to the choice of the posterior probability as a quantity of interest and of the predictive probability criterion) actually coincides with the Bayesian power, defined in Spiegelhalter et al. (2004). Hence, in the next chapter we focus on this case describing in details the power-based methods for SSD, that are the most commonly used in the applications. In order to highlight the main drawbacks of the classical power, we illustrate an example that motivates the introduction of a methodology based on what we name Predictive Bayesian Power. When adopting the two-priors approach this is a generalized power function that simultaneously allows one to exploit pre-experimental information and to take into account the uncertainty on the design value.

Moreover in the second part of the thesis we extend the proposed methodology in two main directions.

- First of all, in Chapter 3, we introduce a robust version of the SSD predictive criteria, in order to address the issue of sensitivity to the elicitation of a single prior distribution, which is one of the most common criticism toward the Bayesian perspective. In particular, the impact of a single prior specification on the optimal sample sizes can be evaluated by considering suitable classes of distributions, such as the ε-contamination classes.

- Secondly, the predictive approach is generalized to a setting in which several sources of pre-experimental information are available. A very straightforward way to take into account the initial information derived from each source is to combine the corresponding prior distributions using a mixture with conveniently chosen weights (see Chapter 4).

Finally, as already mentioned, in Chapter 5 we specifically adapt the proposed SSD criteria to the setting of an equivalence study, in which the purpose is showing that the difference between two competing treatments is negligible.

Chapter 2

Power-based Sample Size Determination

2.1 Introduction and motivations

In describing the general framework of a Bayesian predictive approach for SSD, in Section 1.2.3 we have already noticed that if we choose as a posterior quantity of interest the probability $P_{\pi_A}(\theta > \delta|\mathbf{y_n})$ and if we adopt the predictive probability criterion with coincident π_A and π_D, then the predictive quantity p_n actually reduces to the Bayesian Power defined in Spiegelhalter et al. (2004). This is a particularly relevant case in the context of clinical trials in which the power-based methods for SSD are widely used. Hence, we suggest that this formulation of the power function can be further extended thanks to the two–priors approach (see Gubbiotti & De Santis (2008)), that leads us to define what we name Bayesian Predictive Power. As discussed in the previous chapter, we show how this allows one both to model initial uncertainty on the parameter through the design prior and to exploit pre-experimental information using the analysis prior.

In order to introduce this concept, we start here from a different point of view. First of all we recall the standard power function. Then, drawing on a motivating example, we show how the need of accounting for initial uncertainty leads us to introduce a predictive version of the frequentist power. On the other hand, in order to incorporate prior information into the power formulation, we resort to a Bayesian approach, defining the Bayesian Predictive Power and showing that it can be thought as a "generalized" power function including the others as special cases. Finally, we provide a unifying interpretation for SSD methods based on the power function, highlighting the differences between the classical and the Bayesian perspective, both from a technical and a conceptual point of view.

2.2 Classical and Bayesian power functions

Let us suppose that the objective of the study is inference on a parameter of interest θ. For the sake of simplicity in this chapter we focus on the normal model, namely we assume that $Y_n \sim N(\theta, \frac{\sigma^2}{n})$, where n is the sample size to be determined. As pointed out in Section 1.4 this framework can be adopted not only with normal data but also when a normal approximation applies, for instance when the estimation of the log odds ratio or of the log hazard ratio is of concern. Several examples illustrated in this thesis derive from Spiegelhalter et al. (2004), where a normal approximation is often considered for the parameters of interest.

2.2.1 Conditional Frequentist Power

Let us consider as a parameter of interest θ the unknown effects difference between two alternative treatments, assuming that a positive value of θ favours the new treatment, while a negative value supports the standard one. Hence, the null hypothesis we want to verify is $H_0 : \theta < 0$ against the alternative $H_1 : \theta \geq 0$. Then the power function is defined as the probability of rejecting H_0, conditional to the parameter value θ. We name this function *Conditional Frequentist Power* and we use the notation $\beta_F^C(\theta)$, where the superscript C and the subscript F respectively stand for *conditional* and *frequentist*. In particular, under the normality assumption, we have that

$$\beta_F^C(\theta) = P\left(Y_n > -\frac{1}{\sqrt{n}} z_\alpha \sigma\right) \Phi\left(\frac{\theta\sqrt{n}}{\sigma} + z_\alpha\right) \tag{2.1}$$

where Φ is the cumulative distribution function of the standard normal random variable and z_α is the quantile of a standard normal at level α. Notice that $\beta_F^C(\theta)$ is a function of the parameter θ.

The traditional frequentist SSD criterion suggests to choose the minimum number n that guarantees a given power of the hypothesis test on the mean θ. Hence our objective is to reach a prefixed power η, at a prespecified significance level (for instance $\alpha = 0.05$). In (2.1) we need to fix a design value θ_D that can be interpreted as the target value of the parameter we aim to detect. In other words we are assuming that the sampling distribution of future data Y_n is $f(y_n; \theta_D) = N(y_n | \theta_D, \sigma^2/n)$. Therefore, for a given variance and for a fixed significance level α, the frequentist power conditional to θ_D is an increasing function of n and the optimal sample size is defined as the minimum number of units that guarantees a given power, i.e.

$$n_F^C = \{\min n : \beta_F^C(\theta_D) > \eta\}, \tag{2.2}$$

where the threshold η can be set conventionally for instance at 80%.

This method is widely used in the applications, although it presents two relevant drawbacks, as pointed out in Section 1.1. First of all, as we said before, the optimal sample size noticeably depends on a prefixed design value for the alternative hypothesis. This yields local optimality of the selected sample sizes. Secondly, adopting a frequentist approach, we do not exploit pre-experimental information. Hence the double solution proposed in Section 1.2 is needed. The use of initial information contributes not only to reduce the overall sample size but also allows for more flexibility, reflecting the actual knowledge on the phenomenon before performing the experiment.

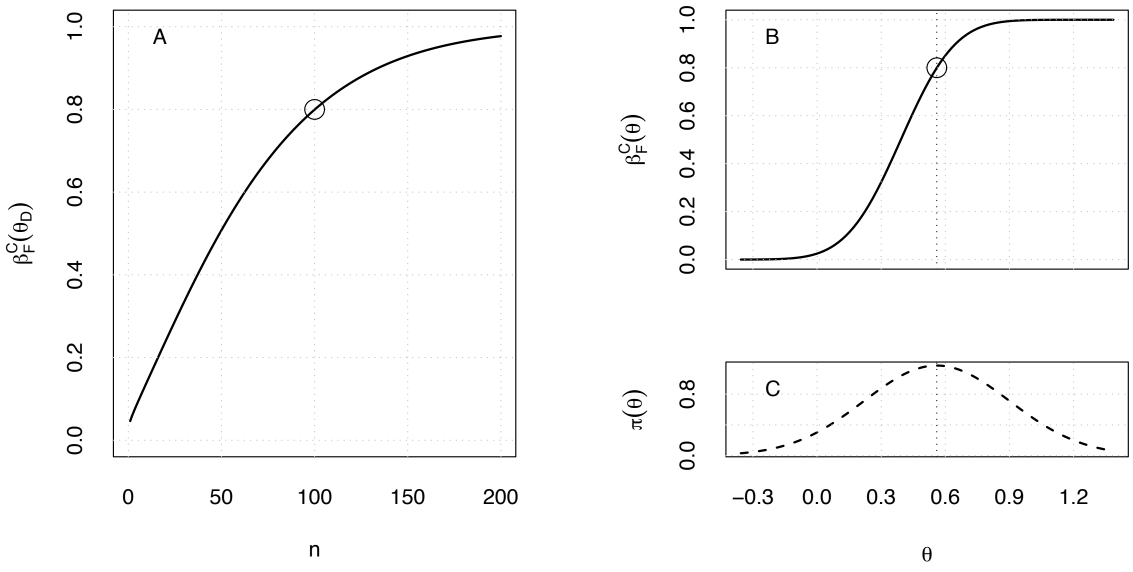
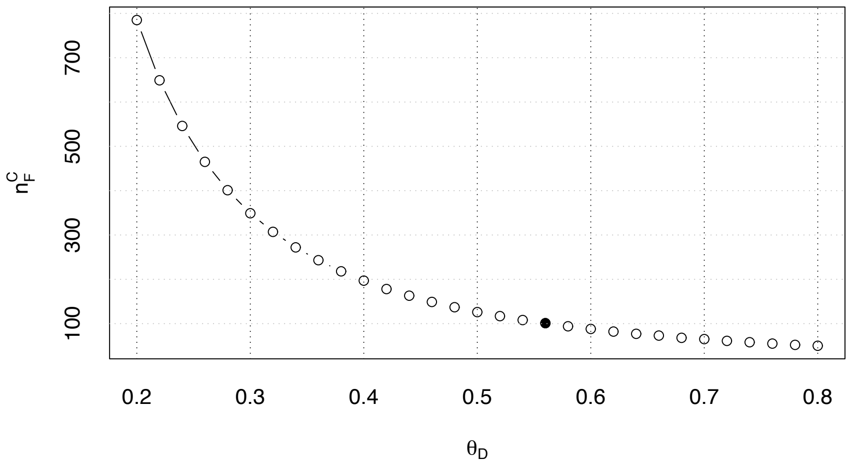


Figure 2.1: **A.** Conditional frequentist power $\beta_F^C(\theta_D)$ with respect to n, where the design value is $\theta_D = 0.56$. The optimal sample size is $n = n_F^C = 100$, corresponding to the required 80% power. **B.** Conditional frequentist power curve $\beta_F^C(\theta)$ with respect to the parameter θ, for a fixed sample size $n = 100$. **C.** Enthusiastic prior for θ: $\pi(\theta)$ is a normal density of mean $\theta_D = 0.56$ and variance $\sigma^2/n_0 = 4/34.5$

Example 1 (*continued*): SSD based on the power function (CANCER) Let us consider again the setting described in Example 1 (page 18) already introduced in Section 1.4 (see of Spiegelhalter et al. (2004)). In panel A of Figure 2.1 for $\theta_D = 0.56$ we obtain the corresponding optimal sample size $n_F^C = 100$. In panel B, however, we highlight the dependence of the frequentist power on θ: for $n = 100$

and $\theta_D = 0.56$ the power is 80% as designed, but it is evident that increasing the design value the power gets higher. In Figure 2.2 we show how the choice of θ_D affects the optimal sample size: the actual values are reported in the table beneath.



θ_D	0.1	0.2	0.3	0.4	0.5	**0.56**	0.6	0.7	0.8
n_F^C	3140	785	349	197	126	**100**	88	65	50

Figure 2.2: Optimal sample sizes n_F^C for several values of θ_D.

In summary, the smaller the effects difference to be detected, the lower the power, the larger the optimal sample size. This intuitive relationship between design value and power clearly shows how crucial the choice of θ_D is, when SSD is of concern. It is then natural to consider a predictive approach that takes into account uncertainty on the design value, as we illustrate in Section 2.2.2. A second remark: the frequentist approach completely ignores possibly available prior information, even in the presence of results from previous studies. For instance, an enthusiastic opinion about the benefit of the new treatment can be expressed by a normal prior density $\pi(\theta) = N(\theta|\theta_D, \sigma^2/n_0)$ centered on a positive value of θ, for example $\theta_D = 0.56$. Then, assuming a remote chance of negative values for θ, for instance a 5% prior probability that $\theta < 0$, we get $\sigma^2/n_0 = 4/34.5$, where n_0 is the so called prior sample size (see Spiegelhalter et al. (2004) for further details). Hence, superimposing the prior $\pi(\theta)$ on the power curve provides a rough indication on the plausibility of the values of the parameter with respect to the corresponding power (see panel C of Figure 2.1). As we said, this procedure just gives an approximate idea; a more formal method is provided by the Bayesian approach that allows one to incorporate the prior $\pi(\theta)$ into the power function and, consequently, in the SSD criterion (see Section 2.2.3).

2.2.2 Predictive Frequentist Power

As shown in the previous section, the conditional frequentist power is strongly related to the chosen design value θ_D, which influences the selection of the sample size. In other words by increasing (or decreasing) the value of θ_D, we reach completely different indications on the optimal sample size for the trial. Instead of considering a single design value we want to take into account the uncertainty around this value in the power function. In this sense, according to the Bayesian approach, we model uncertainty on θ, by specifying a prior probability distribution. Nevertheless suppose for the moment that we do not intend to incorporate prior information in the final analysis, namely we want the conclusions of the study to be entirely classical. This is the idea behind the hybrid classical-Bayesian approach described in Spiegelhalter et al. (2004) and already mentioned in Section 1.1.2. Specifically, we elicit a prior distribution π_D centered on the design value θ_D. Averaging the conditional frequentist power defined in (2.1) with respect to this prior, we obtain the *Predictive Frequentist Power*, that is

$$\beta_F^P(\pi_D) = \int_\Theta \beta_F^C(\theta)\pi_D(\theta)d\theta, \tag{2.3}$$

where the superscript P reminds that it is a *predictive* power function, which corresponds to the unconditional probability of rejecting H_0. The notation also highlights that the predictive power depends on π_D. Again, we assume that π_D is a normal density of mean θ_D and variance denoted by σ^2/n_D. A technical remark: instead of using (2.3), $\beta_F^P(\pi_D)$ can be directly computed as the probability of rejecting the null hypothesis (or equivalently, of getting a significant result) with respect to the marginal distribution of the data. As in Section 1.4, we have

$$m_D(y_n) = N\left(y_n|\theta_D, \sigma^2\left(\frac{1}{n_D} + \frac{1}{n}\right)\right),$$

that is the average of the sampling distribution $f(\cdot;\theta)$ with respect to the prior π_D, according to (1.3). Hence we have

$$\beta_F^P(\pi_D) = \Phi\left(\sqrt{\frac{n_D}{n_D+n}}\left(\frac{\theta_D\sqrt{n}}{\sigma} + z_\alpha\right)\right) \tag{2.4}$$

and it is straightforward to define the following predictive SSD criterion:

$$n_F^P = \{\min n : \beta_F^P(\pi_D) > \eta\} \tag{2.5}$$

for a given threshold η.

Example 1 *(continued)* : SSD based on the power function (CANCER) Let us go back to the example presented in the previous Section. Let us specify for instance the prior $\pi_D(\theta) = N(\theta|\theta_D = 0.56, \sigma^2/n_D = 4/34.5 = 0.16)$. For a sample size $n = n_F^C = 100$, the conditional power $\beta_F^C(\theta_D)$ reaches the required level of 80%, as designed, while the predictive power $\beta_F^P(\pi_D)$ declines to 0.66. In this case, in order to obtain the same power level we should increase the number of observations up to $n_F^P = 240$. In general, we have $n_F^P > n_F^C$.

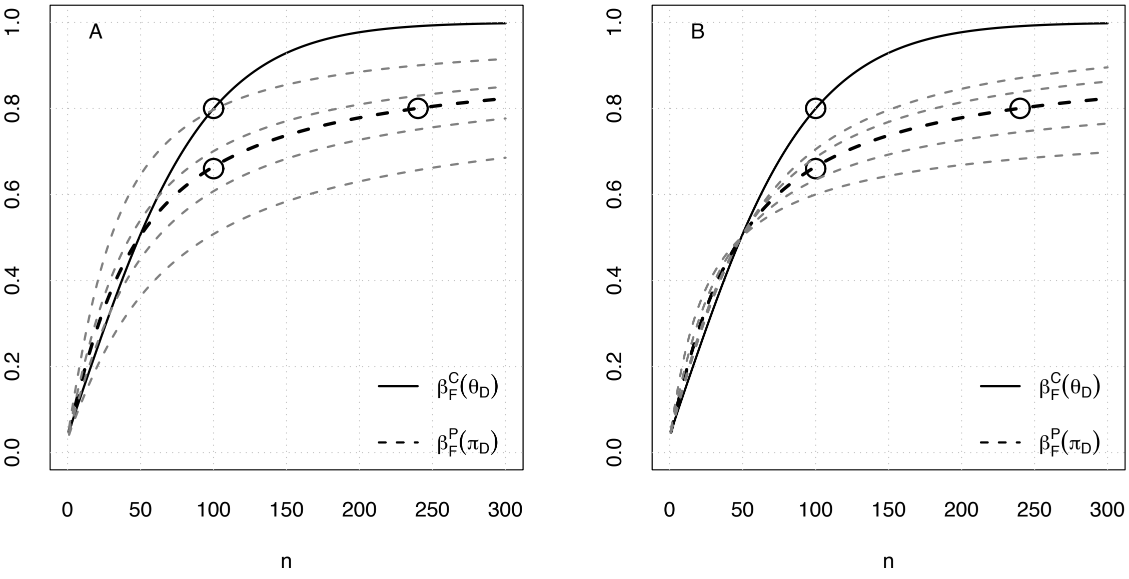


Figure 2.3: $\beta_F^C(\theta_D)$ (solid line) and $\beta_F^P(\pi_D)$ (dashed line) are plotted with respect to n, respectively with design value $\theta_D = 0.56$ and design prior $\pi_D(\theta) = N(\theta|\theta_D = 0.56, \sigma^2/n_D = 4/34.5)$. The dashed gray lines represent $\beta_F^P(\pi_D)$ for different choices of: **A.** the design prior mean $\theta_D = 0.4$, $\theta_D = 0.5$, $\theta_D = 0.6$ and $\theta_D = 0.72$ (from the bottom to the top); **B.** the prior sample size $n_D = 10$, $n_D = 20$, $n_D = 50$ and $n_D = 70$ (from the bottom to the top).

However, from panel A of Figure 2.3 we notice that averaging with respect to the enthusiastic prior π_D slightly raises the power for small values of the sample size, while as n increases the predictive power gets lower than the conditional one. This is even more evident when considering the predictive power curves corresponding to larger prior means, the prior variance being equal. Notice that, in particular, we need to increase the design prior mean to $\theta_D = 0.72$, in order to have $\beta_F^P(\pi_D) = 0.80$ in correspondence to $n = n_F^C = 100$. On the contrary, if we shift the design prior mean towards smaller values – expressing less optimistic opinions on the innovative therapy benefit – we obtain a lower power. In panel B of Figure 2.3 we play on the prior variance, keeping $\theta_D = 0.56$. As expected, for small values of n_D the prior variance increases, that is to say we are actually accounting for more uncertainty, which reduces the predictive power. Viceversa, if we consider larger n_D, the prior π_D gets more informative, raising the predictive power curve. From the comparison of (2.1) and (2.4) it follows that $\beta_F^C(\theta_D)$ is a special case of $\beta_F^P(\pi_D)$: in fact, as $n_D \to \infty$ the prior π_D tends to concentrate on θ_D and $\beta_F^P(\pi_D)$ tends to $\beta_F^C(\theta_D)$. However, as already discussed, for finite n_D we have $\beta_F^P(\pi_D) > \beta_F^C(\theta_D)$, provided that $\beta_F^P(\pi_D) > 0.50$. Finally, notice that if we let $n_D \to 0$, which implies adopting a non informative flat design prior, from (2.4) we have $\beta_F^P(\pi_D) = 0.5$, regardless of the sample size. In other words if we want the predictive SSD criterion in (2.5) to be conclusive, we need to specify a proper design prior.

2.2.3 Bayesian powers

Let us suppose now that in planning the experiment initial information on the treatments difference is available, for example, the results of a previous trial or a pilot study. If we are willing to perform a fully Bayesian analysis, for instance we elicit the prior distribution $\pi_A(\theta) = N(\theta|\theta_A, \sigma^2/n_A)$, where n_A is the prior sample size. Note that the subscript A here stands for *analysis* because we mean this prior to be used in the inferential phase. In the following we consider the general case in which π_A is not necessarily coincident with the π_D that appears in (2.3), as discussed in Section 1.2.2.

Inference is based on the posterior distribution of θ, given the data Y_n. As recalled in Section 1.4, from standard Bayesian analysis it is well known that the posterior is a normal density of parameters given by (1.11) and (1.12). Now, following Spiegelhalter et al. (2004), we say a Bayesian result is *significant* if we have a low posterior chance, say $\alpha = 0.05$, that θ is negative and this happens whenever the following event occurs:

$$Y_n > \frac{-\sqrt{n_A + n}\, z_\alpha \sigma - n_A \theta_A}{n}. \tag{2.6}$$

At this point, according to the choice of the distribution for future data Y_n, we define:

- *Conditional Bayesian Power*:

$$\beta_B^C(\theta_D) = \Phi\left(\frac{\theta_D\sqrt{n}}{\sigma} + \frac{\theta_A n_A}{\sigma\sqrt{n}} + \sqrt{\frac{n_A + n}{n}} z_\alpha\right) \tag{2.7}$$

  if we compute the probability of (2.6) with respect to the sample distribution $f(\cdot;\theta_D)$;

- *Predictive Bayesian Power*:

$$\beta_B^P(\pi_D) = \Phi\left(\frac{1}{\sigma\sqrt{\frac{1}{n_D}+\frac{1}{n}}}\left(\frac{\sqrt{n_A+n}z_\alpha\sigma + n_A\theta_A + n\theta_D}{n}\right)\right) \tag{2.8}$$

if we compute the probability of (2.6) with respect to the marginal distribution $m_D(\cdot)$.

Note that the expression in (2.8) can be further simplified in case we assume $\pi_A = \pi_D$, that is

$$\beta_B^P(\pi_D) = \Phi\left(\frac{\theta_A\sqrt{n_A+n}\sqrt{n_A}}{\sigma\sqrt{n}} + \sqrt{\frac{n_A}{n}}z_\alpha\right) \tag{2.9}$$

where we simply set $\theta_A = \theta_D$ and $n_A = n_D$. This is actually the only case considered in Spiegelhalter et al. (2004); nevertheless, we want to highlight again that two distinct priors may be employed, as pointed out in Section 1.2.2. From the definitions above it is immediate to establish the corresponding SSD criteria, respectively

$$n_B^C = \{\min n : \beta_P^C(\theta_D) > \eta\} \tag{2.10}$$

and

$$n_B^P = \{\min n : \beta_B^P(\pi_D) > \eta\} \tag{2.11}$$

where the threshold η is conventionally equal to 0.80.

In summary, as we show in the example below (see in particular Figure 2.5), we have that

$$n_i^C > n_i^P \text{ for } i = F, B. \tag{2.12}$$

At the same time, adopting an enthusiastic analysis prior, we have that

$$n_F^j > n_B^j \text{ for } j = C, P. \tag{2.13}$$

Example 1 *(continued)* : SSD based on the power function (CANCER) Let us assume for instance that the previous study provides an optimistic indication about the new treatment in terms of log hazard ratio, that can be formalized choosing $\theta_A = \theta_D = 0.56$ and $n_A = n_D = 34.5$. In Figure 2.4 the conditional frequentist and Bayesian powers are represented with respect to n (top panels) and θ (bottom panels). Using the enthusiastic prior π_A, for $n = 100$ we notice an increase in the power up to 0.93. This results in a smaller optimal sample size $n_B^C = 53$ with respect to $n_F^C = 100$. In panel A we also compare the impact of different choices for the prior means on the optimal sample size, being the prior sample size fixed to $n_A = 34.5$.

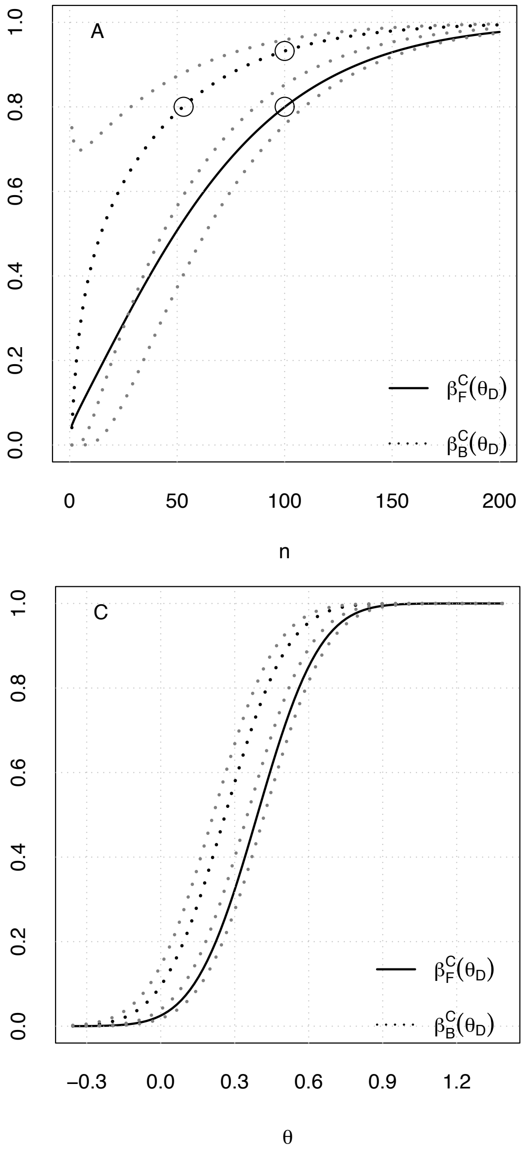


Figure 2.4: Conditional frequentist (solid line) and Bayesian (dotted lines) power curves $\beta_B^C(\theta_D)$, with $\theta_D = 0.56$, are plotted **A.** with respect to n (with $\theta_D = 0.56$) and **C.** with respect to θ (with $n = 100$), for several values of the analysis prior means $\theta_A = 0.1$, $\theta_A = 0.3$, $\theta_A = 0.56$, $\theta_A = 0.7$ (dotted gray lines from right to left), with fixed prior sample size $n_A = 34.5$; **B.** with respect to n (with $\theta_D = 0.56$) and **D.** with respect to θ (with $n = 100$), for several values of the analysis prior sample size $n_A = 0$ (coinciding to $\beta_F^C(\theta_D)$), $n_A = 20$, $n_A = 34.5$, $n_A = 50$ and $n_A = 70$ (dotted gray lines from right to left), with given prior mean $\theta_A = 0.56$.

As expected, the more enthusiastic the prior mean θ_A the higher the power. On the contrary, a prior mean expressing scepticism towards the treatments difference (for instance $\theta_A = 0.1$) leads to a Bayesian power $\beta_B^C(\theta_D)$ uniformly lower than $\beta_F^C(\theta_D)$, the conditional value being equal. In panel B, we proceed in the opposite way: we fix $\theta_A = 0.56$ and plot $\beta_B^C(\theta_D)$ for several values of the prior sample size n_A. Note that the conditional frequentist power is a special case of $\beta_B^C(\theta_D)$ corresponding to $n_A = 0$, i.e. to a flat non informative prior. Then, considering increasing values of n_A we observe at each step a raise in the Bayesian power curve, since the enthusiastic prior gets more and more informative. Similar remarks can be drawn from panel C and panel D, where $\beta_B^C(\theta)$ is plotted with respect to θ, for fixed $n = 100$.

Let us focus now on the predictive Bayesian power curve. As discussed in Section 2.2.2, taking into account the uncertainty on the parameter in the design phase we obtain a lower power, since we are averaging the power function with respect to the design prior. This is evident in Figure 2.5 where the conditional Bayesian power (dotted curve) is compared with the predictive one (dashed-dotted curve). The plot clearly summarizes what we pointed out in (2.12) and (2.13). Furthermore in Figure 2.6 we represent the predictive Bayesian power $\beta_B^P(\pi_D)$ with respect to the sample size n. Playing on the design prior parameters, we reach similar conclusions to the ones in Section 2.2.2: increasing the uncertainty on the design value (i.e. decreasing the prior sample size n_D), the power curve raises (see panel A). The same happens if we choose larger and larger design prior means.

2.3 Concluding remarks

Finally it is interesting to remark that $\beta_P^B(\pi_D)$ can be actually considered as a generalized power function including the other power functions as special cases, as summarized in Table 2.3. Using $\beta_P^B(\pi_D)$ we model both the prior information and the uncertainty on the design value, that can be formalized using – eventually different – prior distributions. Now, if n_D tends to be infinitely large, the design prior tends to a point-mass, i.e. a distribution that assigns probability 1 to the single point θ_D, and we get $\beta_C^B(\theta_D)$ conditional to the design value θ_D. On the other hand if we keep n_D finite and we let n_A go to 0, the analysis prior degenerates in a flat non-informative prior, so that we obtain $\beta_P^F(\pi_D)$, the predictive frequentist power. The conditional frequentist power comes out when we let simultaneously $n_D \to \infty$ and $n_A \to 0$. This means that both design uncertainty and prior information are ignored. Figure 2.5 allows us to compare the behaviour of the four power functions as the sample size n increases.

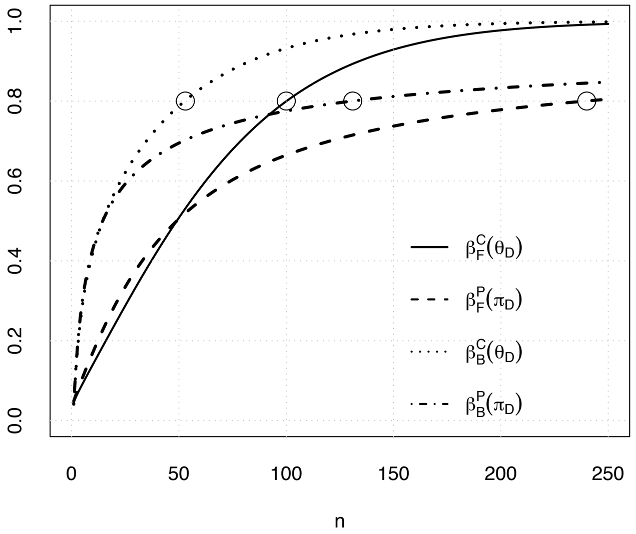


Figure 2.5: $\beta_C^F(\theta_D)$ (continuous line), $\beta_P^F(\pi_D)$ (dashed line), $\beta_C^B(\theta_D)$ (dotted line) and $\beta_P^B(\pi_D)$ (dashed-dotted line) are plotted with respect to the sample size n. The conditional value is $\theta_D = 0.56$; the prior parameters are $\theta_D = \theta_A = 0.56$, $n_D = n_A = 34.5$. The resulting optimal sample sizes are: $n_F^C = 100$, $n_F^P = 240$, $n_B^C = 53$, $n_B^P = 131$

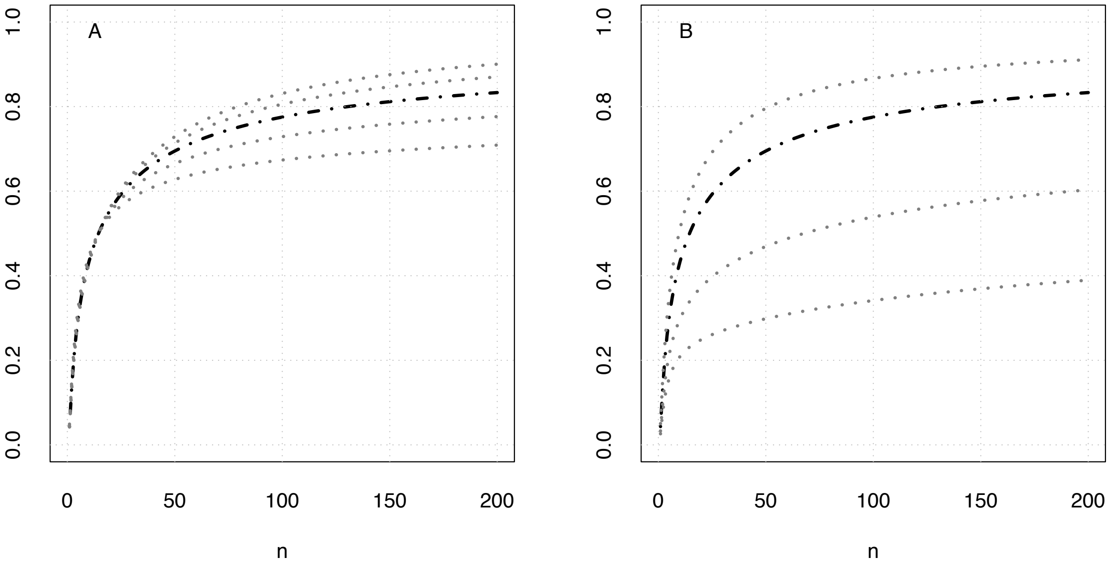


Figure 2.6: The Bayesian (dotted lines) predictive power curve $\beta_B^P(\pi_D)$, is plotted for different choices of the design prior π_D: **A.** $\theta_D = 0.56$ and from the bottom to the top $n_D = 10$, $n_D = 20$, $n_D = 34.5$, $n_D = 50$ and $n_D = 70$; **B.** $n_D = 34.5$ and from the bottom to the top $\theta_D = 0.1$, $\theta_D = 0.3$, $\theta_D = 0.56$ and $\theta_D = 0.7$.

	modeling uncertainty: design prior π_D	
modeling information: analysis prior π_A	$\mathbf{n_D} \to \infty$ $\mathbf{f}(\cdot;\boldsymbol{\theta}_\mathbf{D})$	$\mathbf{n_D} < \infty$ $\mathbf{m}_{\pi_\mathbf{D}}(\cdot)$
non-informative prior: $n_A \to 0$	$\beta_F^C(\theta_D)$	$\beta_F^P(\pi_D)$
proper prior: $n_A > 0$	$\beta_B^C(\theta_D)$	$\beta_B^P(\pi_D)$

Table 2.1: Classification of the power functions according to the use of prior information and the account for uncertainty on the design value: the predictive Bayesian power function can be thought as a general power function including the other three as special cases.

In summary, in this chapter we have first presented the most common SSD criterion, based on the frequentist conditional power. Nevertheless we have argued that this criterion is not flexible enough. In particular, we have underlined that conditioning with respect to a fixed design value, one takes no notice of uncertainty on this value. This consideration has led us to introduce a predictive approach that is able to incorporate uncertainty through a prior distribution, which guarantees a more careful choice of the optimal sample size. On the other hand it is also convenient to exploit eventual prior information directly in the SSD procedure. In fact it is possible to resort to a Bayesian approach that incorporates prior information in the power function. This allows one to take advantage of previous results or opinions of experts about the experiment and in a sense to “spare” sample units in the actual trial.

Chapter 3

Robust Sample Size Determination

3.1 Introduction and motivations

The use of a specific prior distribution for posterior analysis has always been a critical point of Bayesian statistics. This is due to the high degree of subjectivism intrinsic to the selection of one particular distribution. An attempt to address this objection is represented by the *robust Bayesian approach* that:

(i) replaces a single prior with a class of distributions that gives a more flexible and realistic representation of pre-experimental knowledge;

(ii) studies how posterior inference changes as the prior varies over the class.

The idea is simple: if the range of posterior quantities of interest is small, the differences between the various priors in the class are irrelevant and it is possible to 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), Wasserman (1992). In Berger et al. (2000) the Authors present an overview of the robust Bayesian approach, discussing the different possible approaches. First of all they highlight that the issue of robustness with respect to the prior distribution derives from the practical impossibility of eliciting a unique distribution. Furthermore they extend an analogous approach to the other elements involved in a Bayesian analysis, such as the likelihood and the loss function. In their words "the main goal of Bayesian robustness is to quantify and interpret the uncertainty induced by partial knowledge of one of the three elements in the analysis". Applications of robust Bayesian analysis to clinical trials are in Greenhouse & Wasserman (1995, 1996), Carlin & Sargent (1996), Sargent & Carlin (1996). Noting that many medical and epidemiological professionals cite their distaste for informative priors as a prime reason for their ongoing aversion to Bayesian methods, Carlin & Perez (2000) try to address these concerns by investigating Bayesian robustness in some practical applications to clinical trials.

In this Chapter, we apply the robust Bayesian philosophy to the SSD problem illustrated in the previous chapter, as proposed in DasGupta & Mukhopadhyay (1994), De Santis (2006), Brutti & De Santis (2008) and Brutti et al. (2008*b*). Our main goal is the introduction of robust SSD criteria that take into account deviations from an elicited base analysis prior distribution for the unknown parameter. For this reason, we replace a single *base prior* with an entire class of distributions close to it. Then we assume that an experiment is successful if the posterior quantity of interest is sufficiently large for any prior belonging to the chosen class. This is equivalent to check that the lower bound of the posterior quantity with respect to the class of prior distribution, exceeds a given threshold. *Robust sample sizes* are selected by looking at summaries of the predictive distribution of this lower bound. Typically, robust sample sizes are larger than those derived using a single prior. In Brutti et al. (2008*b*) one of the goals is to show the inflate of sample sizes determined 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 the robust one. In these cases we say that single-prior sample sizes are actually robust with respect to the class of priors and that the standard procedure provides adequate sample sizes. In particular, in order to model uncertainty on the base analysis prior we consider classes of ε-contaminated priors, studied for instance in Berger & Berliner (1986) and Sivaganesan & Berger (1989) (see Section 3.2.3). These are mixtures of the base prior with classes of distributions that possess some specific features. Therefore we focus on three relevant classes of contaminant priors: (i) the set of all probability distribution, which is obviously the largest class one can consider, (ii) the class of unimodal distributions and (iii) the class of symmetric unimodal distributions. These 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 chapter is organized as follows. In Section 3.3 and 3.4 we provide expressions for the robust SSD criteria respectively for the normal model and for the binomial model. We also illustrate examples, in order to compare the resulting sample sizes to the non robust ones. Finally in Section 1.3 we analyze the asymptotical behaviour of the predictive summaries involved in the SSD criteria (both robust and non robust) in order to define a reasonable method for the choice of the external thresholds involved in the criteria.

3.2 A robust approach to SSD

3.2.1 Preliminaries

Let us go back to the framework described in Section 1.2.1 and let us suppose that, instead of a single analysis prior we are only able to elicit a class of distributions Γ_A. Specifically, let us assume that we single out a prior π_0 that quantifies pre-trial information on θ, but that we are not completely confident in it. Then in order to avoid the sensitivity due to the specification of a single prior we suggest to replace π_0 with a suitable class of distributions Γ_A "close" to it.

In this way we obtain a robust version of success definition: specifically, we say the trial is *robust-successful* if, for any prior in Γ_A, the chosen posterior quantity of interest $\rho_{\pi_A}(\theta|\mathbf{y_n})$ is larger than γ or, equivalently, if $\inf_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|\mathbf{y_n}) > \gamma$, for a prefixed threshold γ.

3.2.2 Criteria

It is then straightforward to derive the robust versions of *Criterion 1* and *Criterion 2* defined in Section 1.2.3. We simply need to replace $\rho_{\pi_A}(\theta|\mathbf{Y_n})$ with $\inf\limits_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|\mathbf{Y_n})$ in (1.4) and (1.6). In details, we have the following robust criteria.

1. **Robust predictive expectation criterion.** Let

$$e_n^r = \mathbb{E}_{m_D}[\inf_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|\mathbf{Y_n})] \tag{3.1}$$

be the expected value of $\inf\limits_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|\mathbf{Y_n})$ with respect to m_D. Given a threshold η_e, the optimal robust sample size is the number of observations satisfying the following condition:

$$n_{e,r}^* = \min\left\{n \in \mathbb{N} : \ e_n^r > \eta_e\right\}. \tag{3.2}$$

This is the *robust effect-size criterion.*

2. **Robust predictive probability criterion.** Let us consider the *robust predictive power*:

$$p_n^r = \mathbb{P}_{m_D}[R_n^r] = \int_{R_n^r} m_D(y_n)dy_n, \qquad (3.3)$$

where R_n^r is the subset of the sample space containing all the samples which yield a robust-successful experiment:

$$R_n^r = \left\{ y_n : \inf_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|\mathbf{Y_n}) > \gamma \right\}.$$

The robust optimal sample size is the smallest number of observations such that p_n^r is larger than a chosen threshold, $\eta_p \in (0,1)$. In symbols:

$$n^*_{p,r} = \min \{ n \in \mathbb{N} : \ p_n^r > \eta_p \}, \qquad \eta_p \in (0,1). \qquad (3.4)$$

At this point two comments are in order.

(i) As for the choice of the thresholds η_e and η_p, a similar argument to the one of Section 1.3 holds true also when considering the robust criteria defined above. In fact, under a given design scenario, the existence of the optimal robust sample sizes $n^*_{e,r}$ and $n^*_{p,r}$ relies on the choice of η_e and η_p. It is then reasonable to pick these thresholds as prespecified percentages of the maximally achievable value of e_n^r and p_n^r in such a way that the optimization problems defined in (3.2) and (3.4) are actually well-posed. This point is further discussed in Section 3.5 (see also Brutti et al. (2008*b*)), where we study the asymptotic behaviour of e_n^r and p_n^r.

(ii) The consequence of replacing π_A with Γ_A (which we assume to contain π_A), is that, in general, for any given δ, γ, η_e and η_p, the robust sample size is larger than the single-prior sample size, namely $n^*_{\pi_A} < n^*_{\Gamma_A}$. 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, namely $n^*_{\Gamma_A} < n^*_{\Gamma_{A'}}$. Numerical examples will be discussed in the applications of Section 3.3 for the normal model and of Section 3.4 for the binomial model.

3.2.3 Robust SSD using $\varepsilon-$contamination classes

In the present section, following Brutti et al. (2008*b*), we specifically refer to ε-contamination classes. First of all, a formal definition is needed. An ε-contamination class is a mixture of a base prior π_0 with a suitable class of distribution Q, possessing particular characteristics. In symbols:

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

where $\varepsilon \in (0,1)$ is the degree of contamination and q is a contaminant prior belonging to the class Q. According to the choice of Q, we have different ε-contamination classes. Among the many available, we consider in particular the following three options for the class Q:

- $Q_{All} = \{$all the distributions$\}$;
- $Q_U = \{$unimodal distributions with the same mode θ_0 of $\pi_0\}$;
- $Q_{US} = \{$unimodal and symmetric distributions with the same mode θ_0 of $\pi_0\}$.

The corresponding ε-contamination classes will be denoted respectively as Γ_{US}, Γ_U 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. As we show in the following sections, this determines very large sample sizes even for small amounts of contamination. The classes Q_U and Q_{US} are still analytically feasible but they considerably restrict the set of possible contaminant distributions compared to Q_{All}. In Sivaganesan & Berger (1989) the Authors proved some helpful results for computing the bounds of a posterior quantity $\rho_{\pi_A}(\theta|\mathbf{y_n})$ as the prior varies in Γ_{US}, Γ_U and Γ_{All}. Note that, for any $\mathbf{y_n}$, $\Gamma_{US} \subset \Gamma_U \subset \Gamma_{All}$ implies that

$$\inf_{\pi_A \in \Gamma_{US}} \rho_{\pi_A}(\theta|\mathbf{y_n}) \geq \inf_{\pi_A \in \Gamma_U} \rho_{\pi_A}(\theta|\mathbf{y_n}) \geq \inf_{\pi_A \in \Gamma_{All}} \rho_{\pi_A}(\theta|\mathbf{y_n}).$$

Hence, consistently with remark (ii) of Section 3.2.2, we obtain

$$n^*_{\Gamma_{All}} > n^*_{\Gamma_U} > n^*_{\Gamma_{US}}, \tag{3.5}$$

as illustrated in the application of Section 3.3 and 3.4.

Finally, since in general it is not possible to derive closed-form expressions for e_n^r and p_n^r, we can resort to Monte Carlo approximations. In practice we proceed according to the following steps:

1. we draw a large number M of samples from the predictive distribution of the data m_D, say $\tilde{y}_n(1), ..., \tilde{y}_n(M)$;
2. for each generated value $\tilde{y}_n(j)$, we compute $\inf_{\pi_A \in \Gamma_\varepsilon} \rho_{\pi_A}[\theta|\tilde{y}_n(j)]$;
3. we compute the required predictive summaries e_n^r and p_n^r, respectively as a Monte Carlo mean or as the proportion of sampled values exceeding the prefixed threshold γ.

Notice that in step 2 we exploit the results shown by Sivaganesan & Berger (1989) that we resume in the following section.

3.2.4 Bounds of the posterior quantity

As mentioned at the end of the previous section, in order to obtain the bounds of the posterior quantity $\rho_{\pi_A}(\theta|\mathbf{y_n})$, when the analysis prior varies in an ε-contamination class, it is possible to resort to the results of Sivaganesan & Berger (1989). Here we recall the main points of the paper that will be helpful in Section 3.3 and in Section 3.4 where we provide explicit expressions of these bounds with regard to the normal model and the binomial model respectively.

First of all we denote by $m_0(\mathbf{y_n})$ the marginal density of the data induced by the base prior π_0 and we define the following quantities, to be used below:

$$a = (1-\varepsilon)m_0(\mathbf{y_n}) \qquad \text{and} \qquad a_0 = a\rho_{\pi_0}(\theta|\mathbf{y_n}), \tag{3.6}$$

where $\rho_{\pi_0}(\theta|\mathbf{y_n})$ is a posterior summary derived with respect to the base prior π_0. Recall that we focus on a transformation of the parameter of interest $g(\theta)$, that results in a corresponding quantity of interest, according to (1.2) (see also Table 1.1). Let us derive the bounds of $\rho_{\pi_A}(\theta|\mathbf{y_n})$ for the three classes Γ_{All}, Γ_U and Γ_{US}.

Arbitrary contaminations

First of all, for arbitrary contaminations we can distinguish the following two cases:

a. the bounds of the posterior expectation can be obtained computing the inferior and superior extremes of the following expression with respect to θ

$$K_{All} = \frac{a_0 + \varepsilon g(\theta) f_n(y_n;\theta)}{a + \varepsilon f_n(y_n;\theta)} \tag{3.7}$$

b. the bounds of the posterior probability of a set H are respectively:

$$\inf_\theta K_{All} = \frac{a_0}{a + \varepsilon \max\limits_{\theta\in H^c} f_n(y_n;\theta)} \qquad \text{and} \qquad \sup_\theta K_{All} = \frac{a_0 + \varepsilon \max\limits_{\theta\in H} f_n(y_n;\theta)}{a + \varepsilon \max\limits_{\theta\in H} f_n(y_n;\theta)}. \tag{3.8}$$

Unimodal and unimodal symmetric contaminations

As for unimodal or symmetrical unimodal contaminations, in Sivaganesan & Berger (1989) it is shown that the computations are significantly simplified thanks to the alternative representation of a unimodal or symmetrical unimodal distribution as a mixture of uniform distributions. Hence the optimization over Γ_U or Γ_{US} is proved to be equivalent to the optimization over a restriction of these classes, respectively

$$\begin{aligned}
\Gamma_1 &= \{\pi = (1-\varepsilon)\pi_0 + \varepsilon q : q \in U(\theta_0, \theta_0 + z) \text{ or } U(\theta_0 - z, \theta_0) \text{ for some } z > 0\} \subset \Gamma_U,\\
\Gamma_2 &= \{\pi = (1-\varepsilon)\pi_0 + \varepsilon q : q \in U(\theta_0 - z, \theta_0 + z) \text{ for some } z > 0\} \subset \Gamma_{US},
\end{aligned}$$

where $U(a,b)$ is a uniform density in the interval (a,b) and θ_0 is the mode of the base prior π_0. In this way the problem reduces to an optimization with respect to a single variable z, that varies in opportune intervals. We have therefore the following expression for the bounds of $\rho_{\pi_A}(\theta|\mathbf{y_n})$, as π_A varies in Γ_U and Γ_{US}:

$$\begin{aligned} \inf_{\pi\in\Gamma_j} \frac{a_0+\varepsilon K_g(z)}{a+\varepsilon K_{g_0}(z)} &= \inf_{z} \frac{a_0+\varepsilon K_g(z)}{a+\varepsilon K_{g_0}(z)} \\ \sup_{\pi\in\Gamma_j} \frac{a_0+\varepsilon K_g(z)}{a+\varepsilon K_{g_0}(z)} &= \sup_{z} \frac{a_0+\varepsilon K_g(z)}{a+\varepsilon K_{g_0}(z)} \end{aligned} \tag{3.9}$$

for $j=1,2$ respectively, where the quantities K_g are defined below for the two classes Q_U and Q_{US} and K_{g_0} corresponds to the case $g(\theta)=g_0(\theta)=1$.
Hence, for unimodal contaminations we have:

$$K_{g,U}(z) = \begin{cases} \frac{1}{z}\int_{\theta_0}^{\theta_0+z} g(\theta)f_n(y_n;\theta)d\theta & z\neq 0 \\ \\ g(\theta_0)f_n(y_n;\theta_0) & z=0 \end{cases} \tag{3.10}$$

and $K_{g_0,U}(z)=m_q(s_n)$, where m_q is the marginal distribution computed with respect to $q\in U(\theta_0,\theta_0+z)$ or $U(\theta_0-z,\theta_0)$.
Similarly, for unimodal symmetric contaminations we have:

$$K_{g,US}(z) = \begin{cases} \frac{1}{2z}\int_{\theta_0-z}^{\theta_0+z} g(\theta)f_n(y_n;\theta)d\theta & z>0 \\ \\ g(\theta_0)f_n(y_n;\theta_0) & z=0 \end{cases} \tag{3.11}$$

and $K_{g_0,US}(z)=m_q(s_n)$, where m_q is the marginal distribution computed with respect to $q\in U(\theta_0-z,\theta_0+z)$.

3.3 Results for the normal model

In the present Section we derive explicit expressions for the bounds of $\rho_{\pi_A}(\theta|\mathbf{y_n})$ when the normal model is assumed. As for the expression of $\rho_{\pi_0}(\theta|\mathbf{y_n})$, we refer to Section 1.4. Note that under the normality assumptions the marginal distribution computed with respect to the base prior is

$$m_0(y_n) = N\left(y_n|\theta_0, \sigma^2\left(\frac{1}{n}+\frac{1}{n_0}\right)\right).$$

Hence we can determine a and a_0 from (3.6).

Arbitrary contaminations

This is all we need to compute the quantity K_{All} defined in (3.7), that can be then numerically optimized with respect to θ. As for case b, without loss of generality we restrict ourselves to a set of the kind $H = \{\theta : \theta > \delta\}$. Then it is straightforward to derive the following quantities

$$\max_{\theta \in H} f_n(y_n;\theta) = \phi\left(\frac{\sqrt{n}(\delta - y_n)}{\sigma}\right) I_{(-\infty,\delta)}(y_n) + \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} I_{(\delta,+\infty)}(y_n)$$

and

$$\max_{\theta \in H^c} f_n(y_n;\theta) = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} I_{(-\infty,\delta)}(y_n) + \phi\left(\frac{\sqrt{n}(\delta - y_n)}{\sigma}\right) I_{(\delta,+\infty)}(y_n).$$

to be used in computing the exact bounds of (3.8). This accomplishes the case of arbitrary contaminations.

Unimodal and unimodal symmetric contaminations

When considering unimodal and unimodal symmetric contaminations, in order to obtain (3.9) it is necessary to compute respectively the integrals in (3.10) and (3.11) for the different choices of $g(\cdot)$. With regard to the options summarized in Table 1.1, we are interested in considering g as

1. the function identically equal to 1, i.e. $g(\theta) = g_0(\theta) = 1$;
2. the indicating function of the set H, i.e. $g(\theta) = I_H(\theta)$;
3. the identity function, i.e. $g(\theta) = \theta$.

Correspondingly, we derive the results presented below.

1.

$$K_{g_0,U}(z) = \frac{1}{z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 - y_n)\right)\right],$$

$$K_{g_0,US}(z) = \frac{1}{2z}\left[\Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 + z - y_n)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(\theta_0 - z - y_n)\right)\right].$$

2.

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

for $z < 0$, while for $z > 0$ we have

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

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

Finally for $z = 0$, $K_{g,U}(0) = K_{g,US}(0) = I_{\{\theta_0 > \delta\}} f_n(y_n; \theta_0)$.

3. From standard calculations, using the integral

$$\int_a^b x\phi(x; \mu, v)dx = \mu\left[\Phi\left(\frac{b-\mu}{v}\right) - \Phi\left(\frac{a-\mu}{v}\right)\right] + \frac{v}{\sqrt{2\pi}}\left[e^{((a-\mu)/v)^2/2} - e^{((b-\mu)/v)^2/2}\right]$$

we can derive:

$$\begin{aligned} K_{g,U}(z) &= \frac{1}{z}y_n\left[\Phi\left(\frac{\theta_0 + z - y_n}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\theta_0 - y_n}{\sigma/\sqrt{n}}\right)\right] + \\ &+ \frac{\sigma}{\sqrt{2n\pi}}\left[e^{\frac{1}{2}\left(\frac{(\theta_0 - y_n)\sqrt{n}}{\sigma}\right)^2} - e^{\frac{1}{2}\left(\frac{(\theta_0 + z - y_n)\sqrt{n}}{\sigma}\right)^2}\right] \end{aligned}$$

and

$$\begin{aligned} K_{g,US}(z) &= \frac{1}{2z}y_n\left[\Phi\left(\frac{\theta_0 + z - y_n}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\theta_0 - z - y_n}{\sigma/\sqrt{n}}\right)\right] + \\ &+ \frac{\sigma}{\sqrt{2n\pi}}\left[e^{\frac{1}{2}\left(\frac{(\theta_0 - z - y_n)\sqrt{n}}{\sigma}\right)^2} - e^{\frac{1}{2}\left(\frac{(\theta_0 + z - y_n)\sqrt{n}}{\sigma}\right)^2}\right]. \end{aligned}$$

Example 1 *(continued)*: Bayesian robust SSD (CANCER) We extend here Example 1 (page 18), presented in Section 1.4 (see Spiegelhalter et al. (2004)) with the application of the robust criteria. Specifically we use the analysis prior introduced before as base prior π_0, that is a normal density centered on $\theta_0 = 0$, with prior sample size $n_0 = 9$. Then we consider several $\varepsilon-$contamination classes for this base prior. Moreover the design scenario is the same depicted in Figure 1.1, namely we have $\theta_D = 0.56$, $n_D = 34.5$ as design parameters and we set $\delta = 0.1$.

The contour plot in Figure 3.1 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 $\eta_e = 0.8$ ($n^*_{e,r} = 124$, 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_e, 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 situations. This is a consequence of the content itself of the contamination class which includes many undesirable distributions such as point masses that are far way from the base prior π_0.

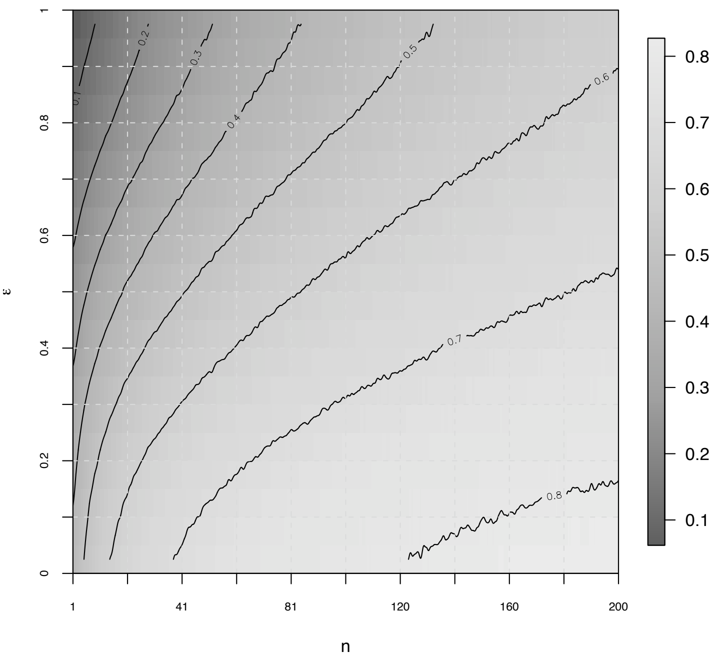


Figure 3.1: Contour plot of e_n^r for Γ_{All} as the sample size n and the contamination parameter ε vary, assuming: $\sigma^2 = 4$, $\theta_0 = 0$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$.

A plausible alternative contamination class is Γ_{US}. In Table 3.1 we summarize 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 3.2. 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.

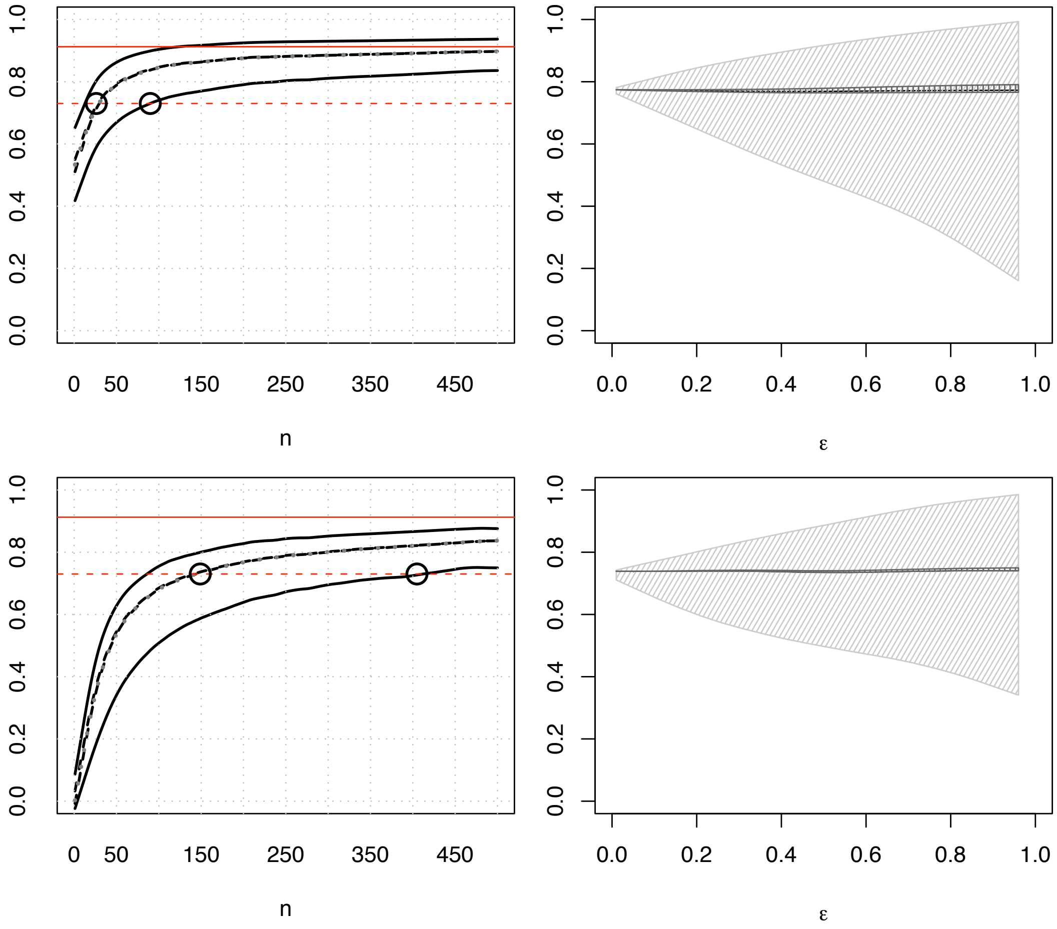


Figure 3.2: e_n^r (top left) and p_n^r (bottom left) for Γ_{All} (two solid lines, representing respectively lower and upper bound) and Γ_{US} (two dashed lines, representing respectively lower and upper bound) as functions of the sample size n (first column, with $\varepsilon = 0.2$)assuming: $\sigma^2 = 4$, $\theta_0 = 0$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$ and $\gamma = 0.9$. The horizontal reference line is set to $\eta_e = \eta_p = 0.73$ ($\beta = 0.8$). In the right panels, predictive summaries of the range of $\rho_{\pi_A}(\theta|\mathbf{y_n})$ (top panel: expectation for $n = n_e^* = 39$; bottom panel: probability for $n = n_p^* = 56$) as the contamination parameter ε varies in $(0, 1)$, respectively for Γ_{All} (light gray area) and for Γ_{US} (dark gray).

In order to observe a wider distance, we 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 3.3, the predictive expectation criterion based on e_n^r leads to more cautious conclusions than the standard criterion e_n.

Moving to the predictive probability criterion (right side of Table 3.1) we see that

		Expectation		**Probability**	
Class	ε	$\theta_0 = 0$	$\theta_0 = 0.29$	$\theta_0 = 0$	$\theta_0 = 0.29$
	0.1	70	49	301 (85)	281 (61)
All	0.2	103	86	408 (142)	381 (120)
	0.3	150	132	477 (208)	456 (190)
	0.1	40	8	152 (49)	137 (6)
US	0.2	42	9	152 (50)	138 (8)
	0.3	43	10	155 (51)	139 (13)
Standard	0.0	39	5	152 (48)	137 (3)

Table 3.1: 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: $\sigma^2 = 4$, $n_0 = 9$, $\theta_D = 0.56$, $n_D = 34.5$, $\delta = 0.1$, $\eta_e = \eta_p = 0.73$ ($\beta = 0.8$), $\gamma = 0.9$ ($\gamma = 0.6$ in brackets), and two different base analysis priors π_0, namely a sceptical one ($\theta_0 = 0$) and an enthusiastic one ($\theta_0 = 0.29$). The line labeled Standard contains the non–robust optimal sample sizes n^*_e and n^*_p (associated to $\varepsilon \equiv 0$).

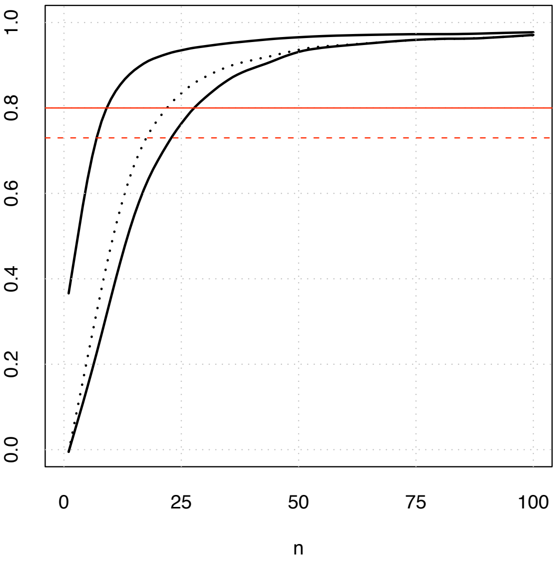


Figure 3.3: e^r_n for Γ_{US} (two solid lines, representing respectively lower and upper bound) as n varies, assuming: $\varepsilon = 0.2$, $\sigma^2 = 4$, $\theta_0 = -1.6$, $n_0 = 9$, $\theta_D = 1.6$, $n_D = 9$, $\delta = 0.1$. The horizontal reference line is set to $\eta_e = 0.73$ ($\beta = 0.8$) , whereas the dotted line corresponds to the standard (non–robust) criterion e_n.

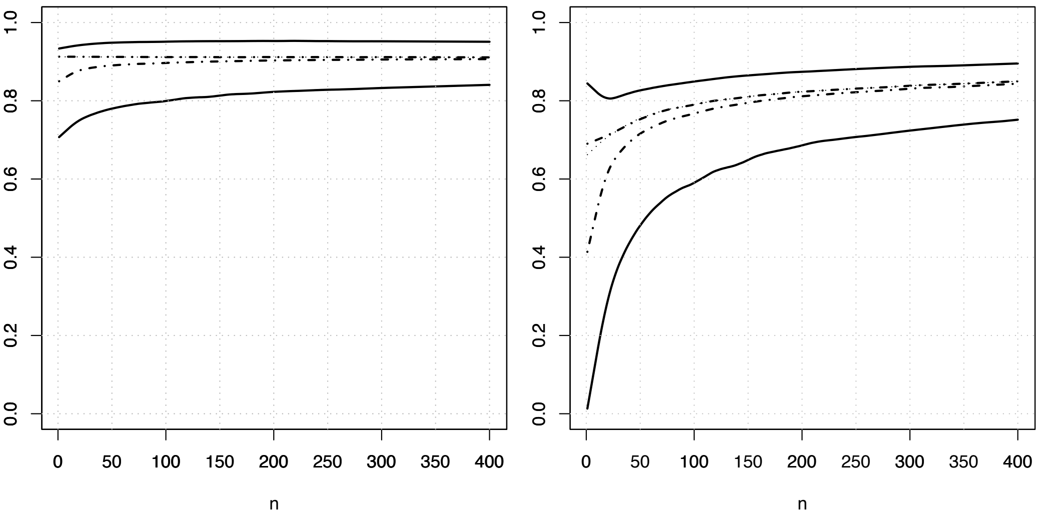


Figure 3.4: e_n^r (left) and p_n^r (right) for Γ_{All} (two solid lines, representing respectively lower and upper bound) and Γ_{US} (two dashed lines, representing respectively lower and upper bound), assuming: $\varepsilon = 0.2$, $\sigma^2 = 4$, $\theta_0 = \theta_D = 0.56$, $n_0 = n_D = 34.5$, $\delta = 0.1$, $\gamma = 0.9$.

all the results are strongly influenced by the value of the parameter γ. As for Γ_{US}, setting $\gamma = 0.6$ leads to optimal sample sizes comparable to those selected by the e_n^r. Increasing the value of γ to 0.9 results in larger values of the optimal sample size, which is coherent with the more strict requirements of the criterion. Finally the optimal sample sizes induced by Γ_{All} are uniformly larger than before because of the higher sensitivity of this criterion to the presence of extreme distributions in the contamination class. Furthermore, if we assume a smaller θ_D, i.e. a smaller true treatment difference, the required sample sizes are even larger. For example, in case we set $\theta_D = 0.29$ corresponding to a hazard ratio of 75%, the predictive probability criterion yields an optimal sample size of about 430 units with $\gamma = 0.9$. Again, the contamination with unimodal symmetric distribution gives comparable results, while the Γ_{All} optimal sample size reaches the unfeasible value of 1267 subjects already for $\varepsilon = 0.1$.

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 3.1 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 3.4). 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 beliefs represented by the base analysis prior.

3.4 Results for the binomial model

In this Section we refer to the binomial model. In particular, following the same scheme of Section 1.5, we consider: (i) the one-sample setting both for the probability of success and the log odds as parameters of interest (Section 3.4.1) (ii) the two–samples setting where the log odds ratio is chosen as a measure of comparison between two treatments effects (Section 3.4.2). Notice again that these cases result from the different specification of the function $g(\cdot)$, as summarized in Table 1.1.

3.4.1 One sample

First of all, let us assume that the base prior distribution for θ is a beta density of parameters (α_0,β_0). The corresponding expression of $\rho_{\pi_0}(\theta|\mathbf{y_n})$ is given in Section 1.5.1. Hence the parameters of the betabinomial marginal distribution computed with respect to the base prior are (α_0, β_0, n). In this setting we derive the bounds of $\rho_{\pi_A}(\theta|\mathbf{y_n})$ for Γ_{All}, Γ_U and Γ_{US}, according to the results of Section 3.2.4.

Arbitrary contaminations

Using (3.6) and (3.7) we obtain the expression to be numerically optimized with respect to θ. As for (3.8), notice that over a set $H = \{\theta : \theta > \delta\}$ (respectively H^c), the maximum of the binomial likelihood $f_n(s_n; \theta)$, considered as a function of θ, depends on the location of the threshold δ with respect to $\hat{\theta} = \frac{s_n}{n}$, that is the maximum likelihood estimate for each couple of values (s_n, n). Hence we have:

$$\max_{\theta \in H} f_n(s_n; \theta) = \left(\begin{array}{c} n \\ s_n \end{array} \right) \hat{\theta}^{s_n} (1 - \hat{\theta})^{n - s_n} I_{\{\hat{\theta} > \delta\}}(y_n) + \left(\begin{array}{c} n \\ s_n \end{array} \right) \delta^{s_n} (1 - \delta)^{n - s_n} I_{\{\hat{\theta} \leq \delta\}}(y_n)$$

and, conversely,

$$\max_{\theta \in H^c} f_n(s_n; \theta) = \left(\begin{array}{c} n \\ s_n \end{array} \right) \hat{\theta}^{s_n} (1 - \hat{\theta})^{n - s_n} I_{\{\hat{\theta} \leq \delta\}}(y_n) + \left(\begin{array}{c} n \\ s_n \end{array} \right) \delta^{s_n} (1 - \delta)^{n - s_n} I_{\{\hat{\theta} > \delta\}}(y_n).$$

Unimodal and unimodal symmetric contaminations

When contaminating the base prior with unimodal and unimodal symmetric distributions, we only need to compute (3.10) and (3.11) for the different choices of $g(\cdot)$, listed in Section 3.3. In details, we have the following results.

1. When $g(\theta) = g_0(\theta) = 1$, for unimodal contaminations, given that $\theta_0 \in (0, 1)$, $z \neq 0$ implies $-\theta_0 \leq z \leq 1 - \theta_0$. Hence we have

$$K_{g_0,U}(z) = \frac{1}{z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\theta_0)\}.$$

On the other hand, for unimodal symmetric contaminations, when $z > 0$ we have

$$K_{g_0,US}(z) = \frac{1}{2z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\theta_0 - z).$$

Finally for $z = 0$

$$K_{g_0,U}(0) = K_{g_0,US}(0) = \left(\begin{array}{c} n \\ s_n \end{array}\right)\theta_0^{s_n}(1-\theta_0)^{n-s_n}.$$

2. If we set $g(\theta) = I_H(\theta)$, for unimodal contaminations, we distinguish the case in which $z > 0$

$$K_{g,U}(z) = \begin{cases} \frac{1}{z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\theta_0)\} & \delta < \theta_0 \\ \\ \frac{1}{z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\delta)\} & \theta_0 < \delta < \theta_0 + z \\ \\ 0 & \delta > \theta_0 + z \end{cases},$$

from the case in which $z < 0$

$$K_{g,U}(z) = \begin{cases} \frac{1}{z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0) - F_{B(s_n+1,n-s_n+1)}(\theta_0 + z)\} & \delta < \theta_0 + z \\ \\ \frac{1}{z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0) - F_{B(s_n+1,n-s_n+1)}(\delta)\} & \theta_0 + z < \delta < \theta_0 \\ \\ 0 & \delta > \theta_0 \end{cases}.$$

For unimodal symmetric contaminations with positive z we have

$$K_{g,US}(z) = \begin{cases} \frac{1}{2z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\theta_0 - z)\} & \delta < \theta_0 - z \\ \\ \frac{1}{2z}\frac{1}{n+1}\{F_{B(s_n+1,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+1,n-s_n+1)}(\delta)\} & -z < \delta - \theta_0 < +z \\ \\ 0 & \delta > \theta_0 + z \end{cases}$$

Finally for $z = 0$:

$$K_{g,U}(0) = K_{g,US}(0) = I(\theta_0 > \delta) \left(\begin{array}{c} n \\ s_n \end{array} \right) \theta_0^{s_n} (1 - \theta_0)^{n-s_n}$$

3. When $g(\theta) = \theta$, for Γ_U we have $-\theta_0 \leq z \leq 1 - \theta_0$ and

$$K_{g,U}(z) = \frac{1}{z} \frac{s_n + 1}{(n+2)(n+1)} \{F_{B(s_n+2,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+2,n-s_n+1)}(\theta_0)\}$$

For Γ_{US} we have $0 < z \leq \min(\theta_0, 1 - \theta_0)$ and

$$K_{g,US}(z) = \frac{1}{2z} \frac{(s_n + 1)}{(n+2)(n+1)} \{F_{B(s_n+2,n-s_n+1)}(\theta_0 + z) - F_{B(s_n+2,n-s_n+1)}(\theta_0 - z)\}.$$

Finally for $z = 0$

$$K_{g,U}(z) = K_{g,US}(z) = \left(\begin{array}{c} n \\ s_n \end{array} \right) \theta_0^{s_n+1} (1 - \theta_0)^{n-s_n}.$$

Example 2 *(continued)*: Bayesian robust SSD (DRUG) Let us go back to Example 2 (page 25) of Section 1.5.1. We consider here the robust criteria in order to check the sensitivity of the resulting optimal sample sizes to the prior specification. We choose, for instance, the most sceptical design prior ($\theta_D = 0.6$) and in Figure 3.5 and Figure 3.6 we compare the results obtained using three different ε-contamination classes for several levels of contamination, respectively for the expectation and the probability criteria, both for the posterior expectation and the posterior probability.

Note that for each choice of the contamination class, the optimal sample sizes sensibly get larger as the contamination level ε increases. Moreover the relationship expressed by (3.5) holds true: using the wider class, Γ_{ALL}, we obtain larger optimal sample sizes than adopting the other two classes.

Log odds scale

Let us consider now the case in which the parameter of interest is the log odds $\psi = \log\left(\frac{\theta}{1-\theta}\right)$, as in described in the second part of Section 1.5.1. Based on the elements specified in Section 3.4.1 the results for arbitrary contaminations are easily derived. Hence we focus here on unimodal and unimodal symmetric contaminations only.

1. For $g(\theta) = g_0(\theta) = 1$ we note that result 1. given in the first part of the present section holds true.

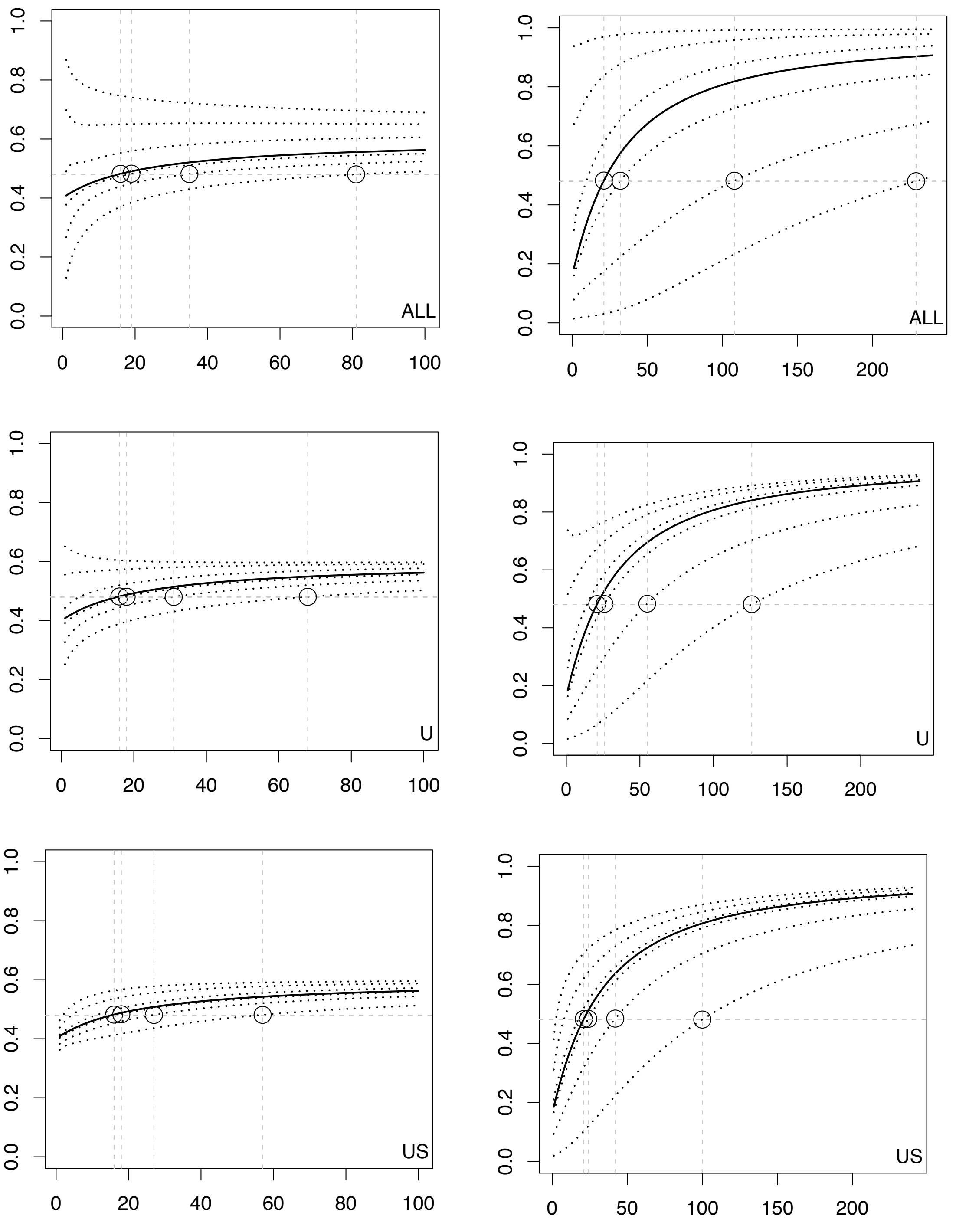


Figure 3.5: Optimal robust sample sizes for Γ_{ALL}, Γ_U and Γ_{US} and for different contamination levels $\varepsilon = 0.1, 0.5, 0.9$, when we consider the predictive expectation of ρ_{π_A}: in the left column we consider the posterior expectation and in the right one the posterior probability with $\delta_1 = 0.5$. Given the threshold $\eta = 0.48$, the resulting optimal sample sizes are reported in Table 3.2.

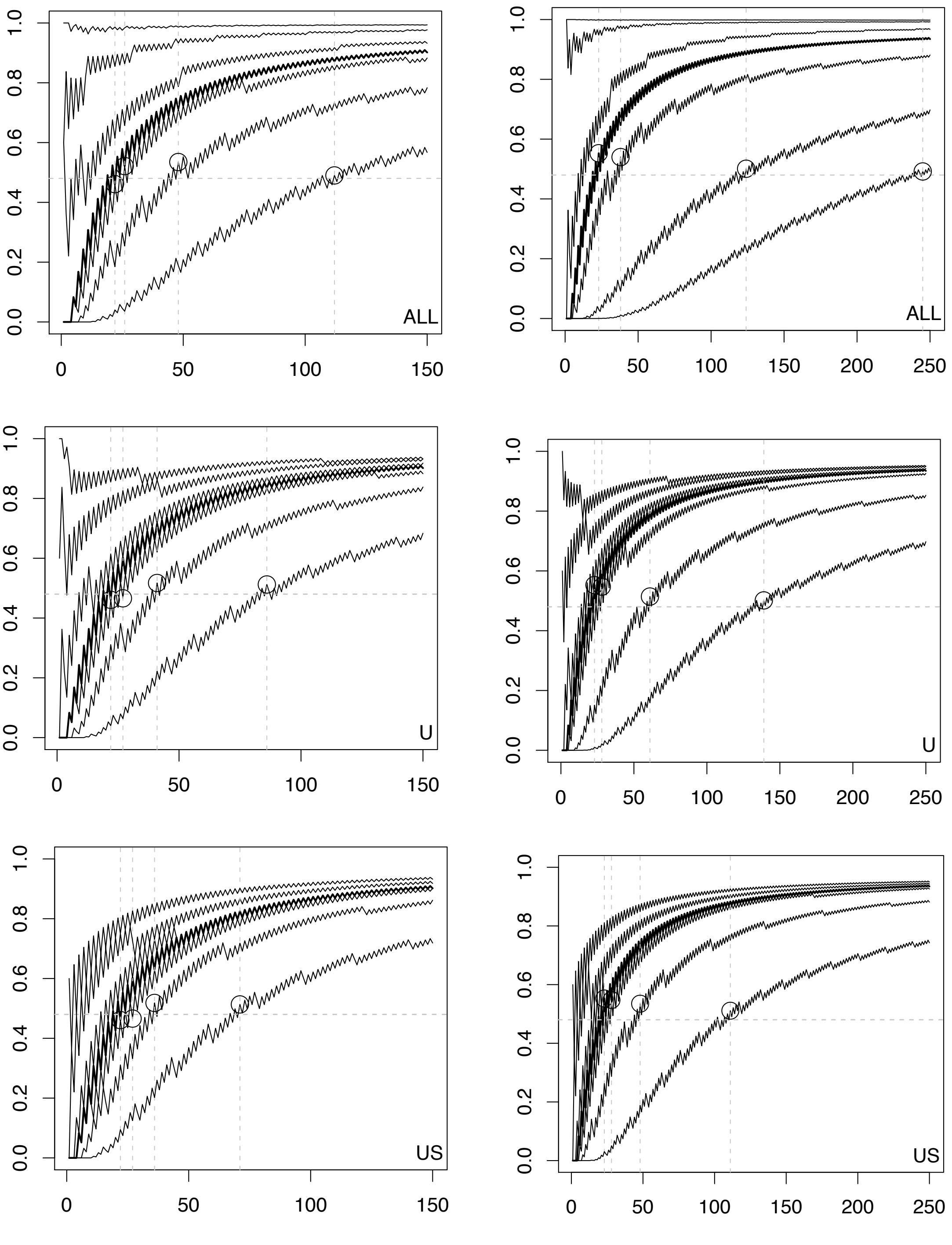


Figure 3.6: Optimal robust sample sizes for Γ_{ALL}, Γ_U and Γ_{US} and for different contamination levels $\varepsilon = 0.1, 0.5, 0.9$, when we consider the predictive probability of ρ_{π_A} (with $\delta_2 = 0.5$): in the left column we consider the posterior expectation and in the right one the posterior probability with $\delta_1 = 0.5$. Given the threshold $\eta = 0.48$, the resulting optimal sample sizes are reported in Table 3.3.

$\rho_{\pi_\mathbf{A}}$	**contamination class**	**contamination level**			
		$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.5$	$\varepsilon=0.9$
E_{π_A}	$\mathbf{\Gamma_{ALL}}$	16	19	35	81
	$\mathbf{\Gamma_{U}}$	16	18	31	68
	$\mathbf{\Gamma_{US}}$	16	18	27	57
P_{π_A}	$\mathbf{\Gamma_{ALL}}$	21	32	108	229
	$\mathbf{\Gamma_{U}}$	21	26	55	126
	$\mathbf{\Gamma_{US}}$	21	24	42	100

Table 3.2: Optimal robust sample sizes for Γ_{ALL}, Γ_U and Γ_{US} and for different contamination levels $\varepsilon = 0.1, 0.5, 0.9$, when we consider the predictive expectation of ρ_{π_A}, for $\delta_1 = 0.5$

$\rho_{\pi_\mathbf{A}}$	**contamination class**	**contamination level**			
		$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=0.5$	$\varepsilon=0.9$
E_{π_A}	$\mathbf{\Gamma_{ALL}}$	22	27	48	112
	$\mathbf{\Gamma_{U}}$	22	27	41	86
	$\mathbf{\Gamma_{US}}$	22	27	36	71
P_{π_A}	$\mathbf{\Gamma_{ALL}}$	23	38	124	245
	$\mathbf{\Gamma_{U}}$	23	28	61	139
	$\mathbf{\Gamma_{US}}$	23	28	48	111

Table 3.3: Optimal robust sample sizes for Γ_{ALL}, Γ_U and Γ_{US} and for different contamination levels $\varepsilon = 0.1, 0.5, 0.9$, for $\delta_1 = 0.5$ and $\delta_2 = 0.5$

2. When we set $g(\theta) = I_H\left(\log\left(\frac{\theta}{1-\theta}\right)\right) = I_H(\psi)$, we have to notice that $H = \{\psi : \psi > \delta'\} = \{\theta : \theta > \frac{e^{\delta'}}{1+e^{\delta'}}\}$. Hence result 2. given in the first part of the present section applies, once we set $\delta = \frac{e^{\delta'}}{1+e^{\delta'}}$.

3. If $g(\theta) = \log\left(\frac{\theta}{1-\theta}\right) = \psi$, for unimodal contaminations we have

$$K_{g,U}(z) = \frac{1}{z}\left(\begin{array}{c} n \\ s_n \end{array}\right)\int_{\theta_0}^{\theta_0+z} \log\left(\frac{\theta}{1-\theta}\right)\theta^{s_n}(1-\theta)^{n-s_n}d\theta,$$

where $z \neq 0$, and for unimodal symmetric contaminations

$$K_{g,US}(z) = \frac{1}{2z}\left(\begin{array}{c} n \\ s_n \end{array}\right)\int_{\theta_0-z}^{\theta_0+z} \log\left(\frac{\theta}{1-\theta}\right)\theta^{s_n}(1-\theta)^{n-s_n}d\theta$$

where $z > 0$ and the above integrals can be computed by Monte Carlo simulation. Finally for $z = 0$

$$K_{g,U}(z) = K_{g,US}(z) = \log\left(\frac{\theta_0}{1-\theta_0}\right)\left(\begin{array}{c} n \\ s_n \end{array}\right)\theta_0^{s_n}(1-\theta_0)^{n-s_n}.$$

3.4.2 Two samples

Finally, let us focus on the log odds ratio φ defined in (1.16). Notice that $\theta = (\theta_1, \theta_2)$ is a vector parameters of two components, but through the transformation $g(\cdot)$ we have $\varphi \in \mathbb{R}$. Hence the results of Sivaganesan & Berger (1989) hold true.

In particular, we need to derive the bounds of a posterior quantity of the kind we defined in Section 1.5.2: the posterior distribution is given by 1.16, assuming a base beta prior for each component θ_i, for $i = 1, 2$, and prior independence between them.

Arbitrary contaminations

First of all, to compute the expressions (3.7) and (3.8), we need to express the likelihood as a function of φ instead of θ. Thanks to the results of Nurminen & Mutanen (1987) and Marshall (1988), already mentioned in De Santis et al. (2004) and in Section 1.5.2, we are able to derive:

$$f_n(y_n; \phi) \propto \exp((s_2+1)\varphi) \int\limits_0^1 \frac{x^{s_1+s_2+1}(1-x)^{n_1-s_1+n_2-s_2+1}}{[1+(\exp(\varphi)-1)x]^{n_2+2}}.$$

Then (3.7) and (3.8) can be optimized numerically.

Unimodal and unimodal symmetric contaminations

As for unimodal and unimodal symmetric contaminations, we need to compute integrals of the kind

$$K_g(z) = \int_l^u \varphi f_n(y_n; \varphi) d\varphi$$

where the extremes of the integral u and l depend on the base prior mode φ_0 and on the variable z. For brevity we omit the details of all the alternative choices of the contamination class and of the function $g(\cdot)$. However in the application in order to compute the integral above we resort to simulation.

Example 3 *(continued)*: Bayesian robust SSD (GREAT)

Let us go back to Example 3 (page 32) introduced in Section 1.5.2. Let us suppose we consider reasonable to contaminate the base prior specified before using the class of unimodal symmetric distributions. Hence, we are in the situation represent in Figure 3.7. Again, notice that the behaviour of e_n (dotted line) and e_n^r (dashed-dotted line) are not very different for a small level of contamination ($\varepsilon = 0.1$ in left panel). Nevertheless, due to the flatness of the considered curves, the resulting sample sizes are quite different: for instance, if we set a threshold $\eta_e = 0.8$ (note that this has to be interpreted on the logOR scale), we have $n_e^* = 132$ and $n_{e,r}^* = 206$. Comparing the three panels of Figure 3.7, we actually show that increasing the level

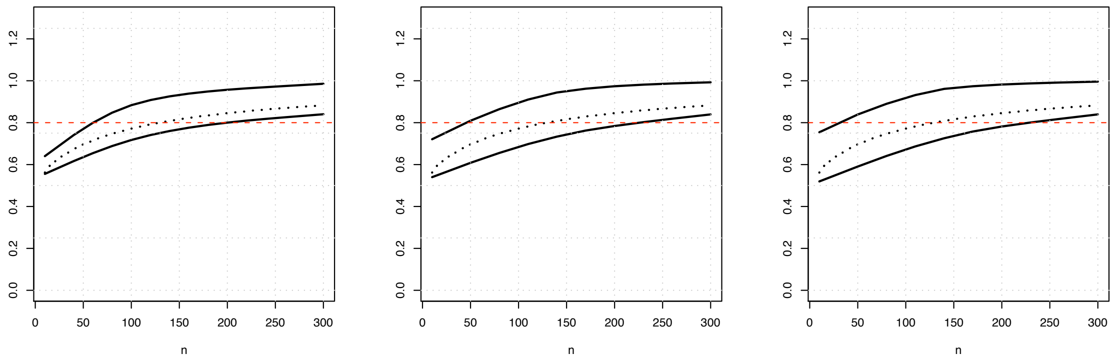


Figure 3.7: e_n (dotted line) and e_n^r (dashed-dotted line) with respect to n, when the posterior quantity of interest is the probability that the log odds ratio exceeds a threshold $\delta = 0$, for $\varepsilon = 0.1$ (left panel), $\varepsilon = 0.5$ (center panel) and $\varepsilon = 0.9$ (right panel).

of contamination does not have dramatic impact on e_n^r. On the other hand, using the class Γ_{All} we obtain again unrealistic results, that would imply unreasonably large sample sizes, even for a small value of ε.

3.5 Asymptotic behaviour of e_n^r and p_n^r

In Section 1.3 we suggested a reasonable criterion for the choice of the thresholds η_e and η_p, based on the study of the asymptotic behaviour of the predictive quantities e_n and p_n involved in (1.5) and (1.7). A similar argument applies when we consider the robust criteria defined in Section 3.2.2. Hence the asymptotic behaviour of e_n^r and p_n^r need to be studied. In practice, in order to obtain the maxima of e_n^r and p_n^r, as the sample size n diverges, it is sufficient to notice that the results proved in Section 1.3 uniformly hold over any class of regular priors like Γ_ϵ. Hence, we have that $\lim_{n\to\infty} p_n^r = \lim_{n\to\infty} p_n = p_\infty$ and $\lim_{n\to\infty} e_n^r = \lim_{n\to\infty} e_n = e_\infty$.

Finally notice that whenever it is not possible to derive a closed-form expression for p_∞ and e_∞, in practice, the limits can be at least numerically approximated. Then, assuming the maximum achievable value as a reference level, the thresholds η_e and η_p can be consequently chosen as described in Section 1.3.

3.6 Concluding remarks

The use of robust techniques in a Bayesian framework allows one to address the critical dependence of the inferential conclusions on the specification of a prior distribution. In the present chapter we deal with this problem in the pre-experimental context, when the size of a trial has to be selected, extending the predictive approach presented in Chapter 1. The main message 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. In order to take into account uncertainty on the base prior, the idea is to replace it with an entire class of priors and to consider the resulting robust sample sizes. In the context of normal and binomial models, we have shown examples in which sample sizes selected using the base prior are very close to robust sample sizes, obtained using the class of unimodal symmetric 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 is interesting whenever the class Γ_{US} is a fairly reasonable representation of prior beliefs on θ. Basically, we now know that sample sizes based on a normal base prior are still adequate under contamination, as long as the contaminated 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 quite larger samples sizes 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 then one can decide to select sample sizes appropriate to the available prior knowledge.

Finally notice that a suitable trade-off is necessary between the level of contamination and the class Q, on the one hand, and the chosen thresholds, 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 is too ambitious (large values of δ, η and γ). In general, the sample size problem turns out to be 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.

Chapter 4

Sample Size Determination and Re-estimation in the presence of multiple sources of information

4.1 Introduction and motivations

In this chapter we adapt the predictive Bayesian approach to determine the size of an experiment, proposed in Chapter 1, to a more complicated setting in which multiple sources of prior information on the unknown parameter of interest θ are available (see Brutti et al. (2008*a*)). In clinical trials it is common, in fact, that pre-experimental information actually derives from distinct historical studies or from the opinions of several expert clinicians. This framework has been recently considered by Gajewski & Mayo (2006) for Phase II clinical trials with binary endpoints. As a prior for θ, the Authors proposed a mixture of conjugate prior distributions, each representing the information derived from every single source, with weights proportional to the degree of pre-experimental "reliability" of each source. Here we propose an extension of the analysis in Gajewski & Mayo (2006) in three main directions. Specifically:

- we consider a predictive approach for pre-posterior sample size computations, following the scheme presented in Chapter 1;

- we adopt the two-priors approach discussed in Section 1.2.2;

- we present results assuming normal endpoints and we illustrate an application (see Section 4.2.3).

The presence of multiple sources of prior information motivates an adjustment of the sample sizes set at the start of the trial after that a portion of experimental outcome has become available. Hence, in addition to the above three points, in Section 4.3 we address the problem of Sample Size Re-estimation (SSRe) based on a first portion of data observed during the ongoing trial. In particular we refer to Wang (2006) where a predictive Bayesian approach is proposed which is based on the expected probability of ending up with a successful trial, given the information provided by the results of the interim analysis. One attractive feature of this methodology in the context described above is that the interim analysis results allow one to update the weights of the mixture components.

4.2 Mixtures of informative priors for SSD

4.2.1 Preliminaries

Let Y_n be an estimator of θ, the unknown quantity of interest in a clinical trial. Let us suppose that K sources of prior knowledge are available for inference on θ, for instance, opinions of K clinicians or data from K historical studies on the experimental medical intervention. The information from each of these sources is formalized in terms of a prior distribution on θ, denoted by $\pi_{A,i}(\theta)$ for $i = 1, \ldots, K$. A standard way to summarize this knowledge is to combine these K priors in a mixture, that is then adopted as analysis prior. Hence we have

$$\pi_A(\theta) = \sum_{i=1}^{K} \omega_{0,i} \pi_{A,i}(\theta), \tag{4.1}$$

where $\omega_{0,i} > 0$ is the prior weight assigned to the i-th component of the mixture, for $i = 1, \ldots, K$, and $\sum_{i=1}^{K} \omega_{0,i} = 1$.
It is straightforward to check that the posterior probability distribution of θ is:

$$\pi_A(\theta|y_n) = \sum_{i=1}^{K} \omega_{1,i}(y_n) \pi_{A,i}(\theta|y_n). \tag{4.2}$$

Each component $\pi_{A,i}$ of the mixture in (4.2) is the posterior probability distribution of θ with respect to the i-th prior according to Bayes theorem

$$\pi_{A,i}(\theta|y_n) = \frac{\pi_{A,i}(\theta) \times f_n(y_n;\theta)}{m_{A,i}(y_n)}$$

where

$$m_{A,i}(y_n) = \int_{\Theta} f(y_n;\theta) \pi_{A,i}(\theta) d\theta$$

is the i-th marginal distribution of the data. Moreover the weight of the i-th posterior distribution can be updated as

$$\omega_{1,i}(y_n) = \frac{\omega_{0,i} m_{A,i}(y_n)}{\sum_{r=1}^{K} \omega_{0,r} m_{A,r}(y_n)}, \qquad i = 1, \dots, K.$$

Let us recover now a similar setting to the one described in Chapter 1. For the sake of simplicity we focus on the posterior probability (see point b in Section 1.2). Then the experiment is defined successful if, for a given $\gamma \in (0, 1)$, we have that:

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

Now we notice that the mixture form (4.1) of the analysis prior, through (4.2), also reflects in the posterior probability of interest defined above. In fact we have:

$$P_{\pi_A}(\theta > \delta | y_n) = \sum_{i=1}^{K} \omega_{1,i}(y_n) P_{\pi_{A,i}}(\theta > \delta | y_n),$$

where it is clear that $P_{\pi_{A,i}}(\theta > \delta | y_n)$ is the posterior probability that θ exceeds δ under the prior $\pi_{A,i}$, for $i = 1, \dots, K$.

4.2.2 Criteria

At this point a similar argument to the one of Section 1.2 applies: before starting the experiment Y_n and, consequently, the posterior quantity of interest $P_{\pi_A}(\theta > \delta | Y_n)$ are random variables. This motivates the need of computing predictive summaries of $P_{\pi_A}(\theta > \delta | Y_n)$ accounting for the randomness of the data in order to establish suitable SSD criteria. First of all we specify a design prior that induces the marginal distribution of the data, defined in (1.3). Then, based on m_D, we compute the requires predictive summary of $P_{\pi_A}(\theta > \delta | Y_n)$; for the sake of brevity we focus here on the predictive expectation only. From (4.1), thanks to the linearity of the expected value, we have that

$$e_n = \mathbb{E}_{m_D}\left[\sum_{i=1}^{K} \omega_{1,i}(Y_n) P_{\pi_{A,i}}(\theta > \delta | Y_n)\right] = \sum_{i=1}^{K} \mathbb{E}_{m_D}\left[\omega_{1,i}(Y_n) P_{\pi_{A,i}}(\theta > \delta | Y_n)\right], \quad (4.3)$$

that is e_n is the sum of the predictive expectations of the terms

$$\omega_{1,i}(Y_n) P_{\pi_{A,i}}(\theta > \delta | Y_n), \qquad i = 1, \dots, K.$$

Then we adopt the Criterion 1 given in (1.5) for the selection of the optimal sample size. In Section 4.2.3 we provide explicit expressions of (4.3) for the normal model when a mixture of conjugate normal distributions is assumed as analysis prior.

4.2.3 Results for the normal model

Assume now that $Y_n|\theta \sim N\left(\theta, \frac{\sigma^2}{n}\right)$ and that each component of the prior is

$$\pi_{A,i}(\theta) = N\left(\theta|\theta_{A,i}, \frac{\sigma^2}{n_{A,i}}\right), \quad i = 1, \ldots, K.$$

In Section 1.4 we remind the standard results on conjugate analysis for the normal model for the posterior mean and variance. For each component we use here the following notation for the posterior mean and the posterior variance:

$$E_{A,i}(\theta|y_n) = \frac{n_{A,i}\theta_{A,i} + ny_n}{n_{A,i} + n} \qquad \text{and} \qquad V_{A,i}(\theta|y_n) = \frac{\sigma^2}{n_{A,i} + n}, \tag{4.4}$$

while we denote by $v_{A,i} = \sigma^2(n_{A,i}^{-1}+n^{-1})$ the variance of the i-th marginal distribution $m_{A,i}$, for $i = 1, \ldots, K$.
Hence, we are able to update the prior weights $\omega_{0,i}$, as follows

$$\omega_{1,i}(y_n) = \frac{\omega_{0,i}\phi\left(\frac{y_n - \theta_{A,i}}{\sqrt{v_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{y_n - \theta_{A,r}}{\sqrt{v_{A,r}}}\right)}.$$

Furthermore, given that

$$P_{A,i}(\theta > \delta|y_n) = 1 - \Phi\left(\frac{\delta - E_{A,i}(\theta|y_n)}{\sqrt{V_{A,i}(\theta|y_n)}}\right),$$

we derive the explicit expression of (4.3) under the normal assumption:

$$e_n = \sum_{i=1}^{K}\mathbb{E}_{m_D}\left\{\frac{\omega_{0,i}\phi\left(\frac{Y_n - \theta_{A,i}}{\sqrt{v_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{Y_n - \theta_{A,r}}{\sqrt{v_{A,r}}}\right)}\left[1 - \Phi\left(\frac{\delta - E_{A,i}(\theta|Y_n)}{\sqrt{V_{A,i}(\theta|Y_n)}}\right)\right]\right\}. \tag{4.5}$$

Finally, we also assume normality for the design prior and, consequently, for the marginal distribution (see Section 1.4). In order to compute the expected value in (4.5) with respect to m_D we resort to Monte Carlo simulation.

Of course the method for the choice of the threshold η_e, discussed in Section 1.3, holds true. Hence, in order to tune η_e we start by evaluating the suprema of e_n (an increasing function of n) for given δ and design prior, e_∞. Then, we take η_e as a prespecified percentage $\beta \in (0, 1)$ of e_∞, so as to ensure the existence of the optimal sample size n_e^*. Therefore we need first of all to discuss the asymptotic behaviour of e_n. In next section we show that e_n converges to a quantity e_∞ that can be computed via a Monte Carlo approximation.

4.2.4 Asymptotic behaviour of e_n

In order to apply the criterion proposed in Section 1.3 for the choice of threshold η_e, preliminarily we have to study the asymptotic behaviour of (4.5). First of all notice that as $n \to \infty$ we have that:

- the posterior mean of the $i-$th component $E_{A,i}(\theta|Y_n)$ asymptotically behaves as Y_n;
- the posterior variance of the $i-$th component, $V_{A,i}(\theta|Y_n)$, tends to 0 (a.s.);
- the variance of the marginal distribution induced by the $i-$th prior component, $v_{A,i}$, converges to $\sigma^2/n_{A,i}$ (prior variance);
- the sequence of random variables Y_n, with marginal densities m_D, converges to $N\left(\theta_D, \frac{\sigma^2}{n_D}\right)$, whose density is here denoted as m_∞.

Hence, by the dominated convergence theorem, the limit of (4.5) is

$$\lim_{n\to\infty} e_n = \sum_{i=1}^{K} \lim_{n\to\infty} \mathbb{E}_{m_D}\left\{ \frac{\omega_{0,i}\phi\left(\frac{Y_n-\theta_{A,i}}{\sqrt{v_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{Y_n-\theta_{A,r}}{\sqrt{v_{A,r}}}\right)} \left[1-\Phi\left(\frac{\delta - E_{A,i}(\theta|Y_n)}{\sqrt{V_{A,i}(\theta|Y_n)}}\right)\right]\right\} \tag{4.6}$$

$$= \sum_{i=1}^{K}\int_{\mathbb{R}} \lim_{n\to\infty}\left\{ \frac{\omega_{0,i}\phi\left(\frac{y_n-\theta_{A,i}}{\sqrt{v_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{y_n-\theta_{A,r}}{\sqrt{v_{A,r}}}\right)} \left[1-\Phi\left(\frac{\delta - E_{A,i}(\theta|y_n)}{\sqrt{V_{A,i}(\theta|y_n)}}\right)\right]\right\} m_D(y_n)\mathrm{d}y_n.$$

Note that, as $n \to \infty$, the expression in square brackets converges to 1 or 0 according to the sign of the argument of $\Phi(\cdot)$. Moreover, taking into account the limiting distribution of Y_n, each term of the sum can be written as:

$$\int_{\mathbb{R}}\left\{ \frac{\omega_{0,i}\phi\left(\frac{z-\theta_{A,i}}{\sqrt{v_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{0,r}\phi\left(\frac{z-\theta_{A,r}}{\sqrt{v_{A,r}}}\right)} \quad \mathbb{I}_{[\delta,\infty)}(z)\right\}\cdot m_\infty(z)dz,$$

and, consequently, Equation (4.7) reduces to

$$\sum_{i=1}^{K}\mathbb{E}_{m_\infty}\left[\omega_{1,i}(Z)\cdot \mathbb{I}_{[\delta,\infty)},(Z)\right]$$

which is computed via Monte Carlo approximation.

Example 4: Predictive SSD using a mixture of priors derived from previous studies (MAGNESIUM) We revisit an example in Spiegelhalter et al. (2004) where the results of a meta-analysis are reinterpreted according to a Bayesian perspective, in order to show the degree of scepticism necessary to reach an opposite conclusion with respect to the actually observed one. A series of small randomized trials was conducted in order to prove a protective effect of intravenous magnesium sulphate after acute myocardial infarction. These studies culminated in a meta-analysis which showed a highly significant 55% reduction in odds of death. This was confirmed in 1992 by a larger study (LIMIT-2 trial) that demonstrated a 24% reduction in mortality in 2000 patients. All these results suggested an outstanding conclusion: a cheap, safe and simple treatment reduces mortality in a common condition. For this reason, further investigation was recommended. But the massive ISIS-4 trial did not actually show evidence of any benefit: the final result on 58000 patients showed a non significant protective effect of magnesium, also consistent across major subgroups. Here we draw on this framework in order to formalize the situation in which prior knowledge comes from different historical studies.

i	**study**	**magnesium**		**control**		N_i	$\theta_{\mathbf{A,i}}$	$\frac{\sigma}{\sqrt{\mathbf{n_{A,i}}}}$	$\mathbf{n_{A,i}}$
		deaths	patients	deaths	patients				
1	Morton	1	40	2	36	76	−0.65	1.06	3.6
2	Rasmussen	9	135	23	135	270	−1.02	0.41	24.3
3	Smith	2	200	7	200	400	−1.12	0.74	7.4
4	Abraham	1	48	1	46	94	−0.04	1.17	2.9
5	Feldstedt	10	150	8	148	298	0.21	0.48	17.6
6	Shechter	1	59	9	56	114	−2.05	0.9	4.9
7	Ceremuzynsky	1	25	3	23	48	1.03	1.02	3.8
8	LIMIT-2	90	1159	118	1157	2316	−0.3	0.15	187

Table 4.1: Observed results (logOR scale) in 8 studies on the protective effect of magnesium, standard deviation and effective number of events.

We focus on the log odds ratio as parameter of interest θ. In Spiegelhalter et al. (2004) the Authors suggest to estimate θ by $\hat{\theta} = \log\left(\frac{(a+\frac{1}{2})(d+\frac{1}{2})}{(b+\frac{1}{2})(c+\frac{1}{2})}\right) = y_n$, where a and b denote respectively the number of observed events in the control arm and in the treatment arm, with $a + b = n$, and c and d are the respective numbers of patients in the two groups who did not experience any event. The additional terms $1/2$ have the effect of lessening the bias of the estimator and preventing problems with small numbers of events. Furthermore this generally has a negligible effect when the sample size is reasonably large. Adopting Spiegelhalter et al.'s terminology we want to determine the *effective sample size*, that is actually the total number of events n. Then the corresponding statistic Y_n is asymptotically distributed as a normal density of mean $\hat{\theta}$ and variance σ^2/n, where σ is set equal to 2 (see Spiegelhalter et al. (2004) for further details).

We proceed eliciting a conjugate normal prior distribution based on each historical study, assuming the estimated log odds ratios and the corresponding standard deviations summarized in Table 4.1 as the parameters of the normal prior components. The global analysis prior is then given by a mixture of these eight priors, with conveniently chosen weights. The prior components and the corresponding the mixture are represented in the left panels of Figure 4.1 and Figure 4.2 choosing respectively equal weights or weights proportional to each study dimension N_i. Note that, since the parameter of interest is the logOR of magnesium with respect to placebo, negative values on this scale support the idea of a benefit of magnesium administration. Nevertheless in this case we are actually interested in proving that θ is larger than a threshold δ, meaning that magnesium is not effective. This is not the standard situation of a superiority trial, but the methodology described in Section 4.2.1 and in Section 4.2.2 is essentially the same. Alternatively the problem could be reverted, defining the logOR of placebo with respect to magnesium and focusing on $P_{\pi_A}(\theta < \delta)$ as a posterior quantity of interest. At this point we specify a design prior expressing scepticism towards the treatment. A possible choice can be based on the results of ISIS-4 trial: this yields a design prior which is a normal density with mean 0.058 and effective number of events 4319, resulting in a very small variance (0.00092). In the first row of Figure 4.1 the center and the right panel represent the predictive expectation e_n with respect to n, for two different choices of δ, and the optimal sample size is selected in correspondence of a prespecified threshold $\eta = 0.8$. Since the analysis prior strongly supports the hypothesis of a protective effect of magnesium, we would need a sizeable number of events to be able to reach an opposite conclusion (about 1747 for $\delta = 0$). Moreover if we choose $\delta = -0.1$, the goal is less challenging and only 498 events are required.

δ	$\mathbf{n_D}$	$\mathbf{e_\infty}$	η	$\mathbf{n^*}$	
				equal weights	proportional weights
-0.1	4319	1	0.80	498	457
	432	0.95	0.76	509	460
	43	0.70	0.56	198	169
0	4319	0.97	0.78	1747	2294
	432	0.73	0.58	243	661
	43	0.58	0.46	42	183

Table 4.2: Optimal sample sizes for equal or proportional weights with respect to different design priors, choosing $\eta_e = \beta \cdot e_\infty$, with $\beta = 0.80$

Alternatively we can choose for instance prior weights proportional to the actual dimension N_i of each historical study. In this case we obtain the mixture represented

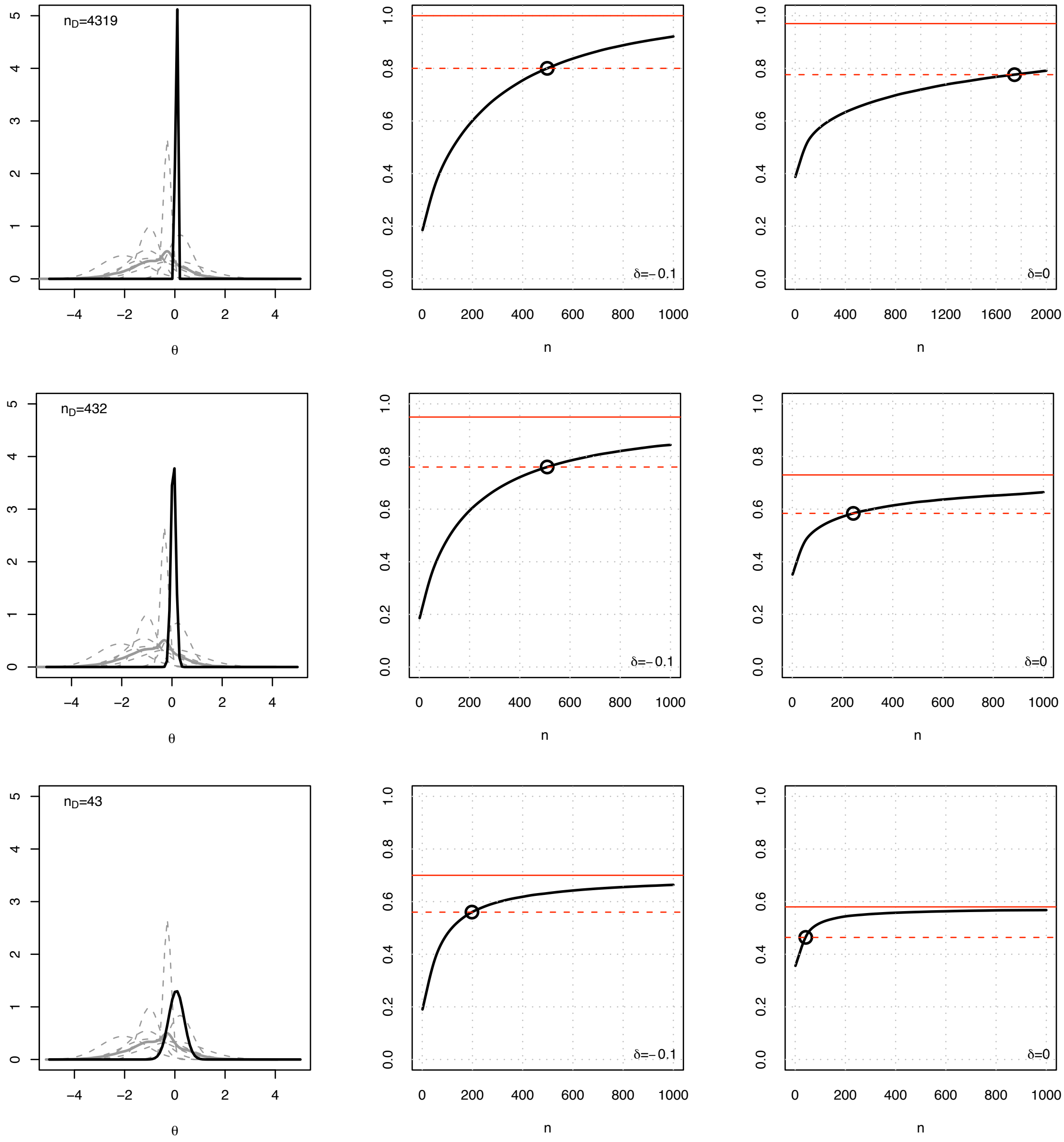


Figure 4.1: (*left panels*) Prior components (dashed gray lines), mixed prior with equal weights (continuous gray line) and design prior (black line), for $n_D = 4319$, $n_D = 432$, $n_D = 43$. Selection of the optimal sample size for $\delta = 0$ (*center panel*) and $\delta = -0.1$ (*right panels*). See Table 4.2

in Figure 4.2 (first row, left panel); then the corresponding optimal sample size is selected. Notice that the prior component of LIMIT-2 trial is highly predominant in the mixed analysis prior ($N_8 = 2316$). This yields larger optimal sample sizes ($n^* = 2294$ for $\delta = 0$ and $n^* = 457$ for $\delta = -0.1$), since the analysis prior is more informative and closer to the design prior. Moreover we considered two less informative design priors with smaller number of events, $n_D = 432$ and $n_D = 43$

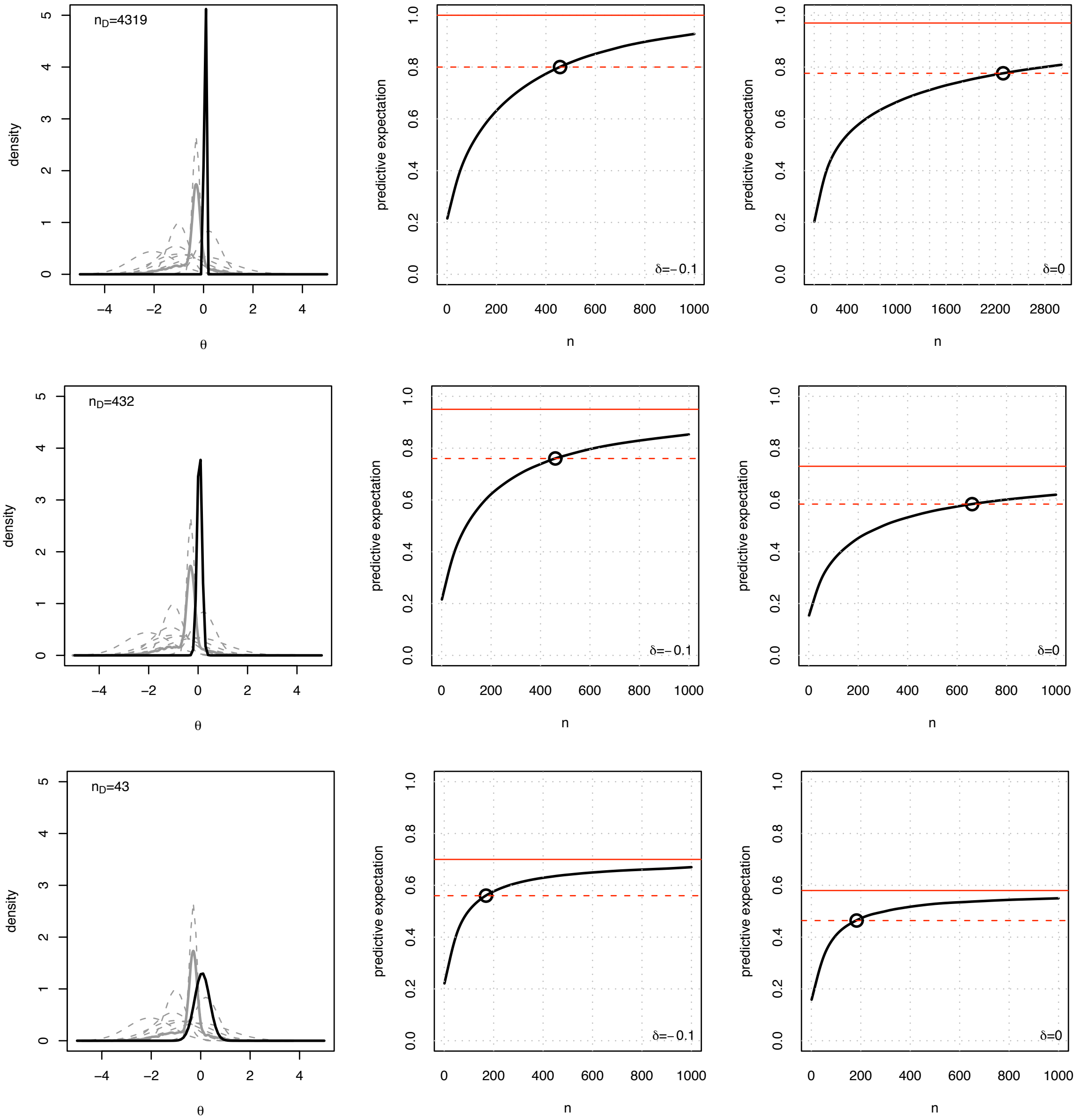


Figure 4.2: (*left panels*) Prior components (dashed gray lines), mixed prior with weights proportional to the dimension of each historical study (solid gray line) and design prior (black line), for $n_D = 4319$, $n_D = 432$, $n_D = 43$. Selection of the optimal sample size for $\delta = 0$ (*center panel*) and $\delta = -0.1$ (*right panels*). See Table 4.2

(see respectively the second row and the third row of Figure 4.1 and 4.2). For each different choice of the design parameters we computed the corresponding e_∞. We set consequently $\eta_e = \beta \cdot e_\infty$, for instance with $\beta = 0.80$, as discussed in Section 1.3 and 4.2.4. The optimal sample sizes are reported in Table 4.2. It is quite evident that the more informative the design prior, the higher the maximum achievable value of

e_n. Notice that, for example, for $n_D = 43$ and $\delta = 0$, e_∞ is equal to 0.58, so if we used a fixed $\eta_e = 0.80$, n_e^* would be undetermined. This supports again the criterion suggested in Section 4.2.4 for the choice of η_e.

4.3 Mixtures of informative priors for SSRe

4.3.1 Preliminaries

A predictive approach is now used for SSRe. Let us assume that, at a given time point, a fraction $n^{(1)}$ of the planned subjects have completed the trial. The objective is then to select the number $n^{(2)}$ of further sample units required to successfully complete the experiment, by exploiting the information contributed by the first $n^{(1)}$ observed events; let us denote by $y_{n^{(1)}}$ the corresponding observed statistic. The idea is to use as initial distribution, at the interim analysis, the posterior density of θ given $y_{n^{(1)}}$, $\pi_A(\theta|y_{n^{(1)}})$. Note that from (4.2) it follows that $\pi_A(\theta|y_{n^{(1)}})$ can be written as a mixture of K different initial priors, whose weights are $\omega_{1,i}(y_{n^{(1)}})$, $i = 1, ..., K$.

In the second part of the trial $n^{(2)}$ events are to be observed, with $n^{(1)} + n^{(2)} = n$. The SSRe problem is to determine $n^{(2)}$. Given the observed value of $y_{n^{(2)}}$ after $n^{(2)}$ events, the posterior distribution can be written as

$$\pi_A(\theta|y_{n^{(1)}}, y_{n^{(2)}}) = \sum_{i=1}^{K} \omega_{2,i}(y_{n^{(2)}}|y_{n^{(1)}})\pi_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}})$$

where

$$\pi_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}}) = \frac{\pi_{A,i}(\theta|y_{n^{(1)}})f_{n^{(2)}}(y_{n^{(2)}};\theta)}{m_{A,i}(y_{n^{(2)}}|y_{n^{(1)}})} \tag{4.7}$$

and where the weights at the interim analysis are

$$\omega_{2,i}(y_{n^{(2)}}|y_{n^{(1)}}) = \frac{\omega_{1,i}(y_{n^{(1)}})m_{A,i}(y_{n^{(2)}}|y_{n^{(1)}})}{\sum\limits_{r=1}^{K} \omega_{1,r}(y_{n^{(1)}})m_{A,r}(y_{n^{(2)}}|y_{n^{(1)}})}, \quad i = 1, ..., K.$$

The posterior predictive distribution of $Y_{n^{(2)}}$ is

$$m_{A,i}(y_{n^{(2)}}|y_{n^{(1)}}) = \int_\Theta f_{n^{(2)}}(y_{n^{(2)}};\theta)\pi_{A,i}(\theta|y_{n^{(1)}})d\theta \tag{4.8}$$

and the posterior quantity of interest is

$$P_{\pi_A}(\theta > \delta|y_{n^{(1)}}, y_{n^{(2)}}) = \sum_{i=1}^{K} \omega_{2,i}(Y_{n^{(2)}}|y_{n^{(1)}})P_{\pi_{A,i}}(\theta > \delta|y_{n^{(1)}}, y_{n^{(2)}}). \tag{4.9}$$

4.3.2 Criteria

Again, note that the posterior quantity in (4.9) is random before $y_{n^{(2)}}$ is observed. Hence, we introduce a predictive criterion to select the optimal additional sample size $n^{(2)*}$:

$$n^{(2)*} = \min\left(n^{(2)} \in \mathbb{N} : e_{n^{(1)},n^{(2)}} > \eta_e\right) \quad \text{for} \quad \eta_e \in (0,1)$$

where

$$\begin{aligned} e_{n^{(1)},n^{(2)}} &= \mathbb{E}_{m_D}\left[P_{\pi_A}(\theta > \delta | y_{n^{(1)}}, Y_{n^{(2)}})\right] = \\ &= \sum_{i=1}^{K} \mathbb{E}_{m_D}\left[\omega_{2,i}(Y_{n^{(2)}}|y_{n^{(1)}}) P_{\pi_{A,i}}(\theta > \delta | y_{n^{(1)}}, Y_{n^{(2)}})\right]. \end{aligned} \tag{4.10}$$

The expected value in (4.10) is now computed with respect to the predictive distribution m_D, induced by the design prior π_D. Note that, at the interim stage, to obtain the predictive density m_D for SSRe we can use either $\pi_D(\theta)$ or $\pi_D(\theta|y_{n^{(1)}})$. In the former case we preserve the initial design goals, expressed by $\pi_D(\theta)$. In the latter we actually adjust design objectives according to the findings of the first part of the experiment. These two alternatives are discussed in the example of Section 4.3.3.

4.3.3 Results for the normal model

In this Section we compute $e_{n^{(1)},n^{(2)}}$, under the normality assumption both for the model and for the prior components of the mixture analysis prior,

Hence, we need to state beforehand the following results. First of all, each posterior component of (4.7) is

$$\pi_i(\theta|y_{n^{(1)}}, y_{n^{(2)}}) = N(\theta|E^{(2)}_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}}), V^{(2)}_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}}))$$

where the posterior mean and variance are respectively

$$E^{(2)}_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}}) = \frac{(n_{A,i} + n_1)E^{(1)}_{A,i}(\theta|y_{n^{(1)}}) + n^{(2)} y_{n^{(2)}}}{n_{A,i} + n^{(1)} + n^{(2)}}$$

and

$$V^{(2)}_{A,i}(\theta|y_{n^{(1)}}, y_{n^{(2)}}) = \frac{\sigma^2}{n_{A,i} + n^{(1)} + n^{(2)}}$$

and $E^{(1)}_{A,i}$ and $V^{(1)}_{A,i}$ are given by (4.4), when the first $n^{(1)}$ observations are considered. Moreover the marginal distribution in (4.8) is a normal density of parameters $(E^{(1)}_{A,i}(\theta|y_{n^{(1)}}), v^{(2)}_{A,i})$, where the variance is given by

$$v^{(2)}_{A,i} = \sigma^2\left(\frac{1}{n_{A,i} + n^{(1)}} + \frac{1}{n^{(2)}}\right)$$

for $i = 1, ..., K$. Note that the expected value of (4.10) is computed with respect to the predictive distribution m_D, which is a normal distribution as well. As discussed in Section 4.3.2, m_D can be alternatively derived using the design prior $\pi_D(\theta)$ or the posterior distribution $\pi_D(\theta|y_{n^{(1)}})$. In the first case we have again the predictive distribution of (1.3), while in the second case we have

$$m_D(y_{n^{(2)}}|y_{n^{(1)}}) = N\left(y_{n^{(2)}}|\frac{\mu_D n_D + n^{(1)} y_{n^{(1)}}}{n_D + n^{(1)}}, \sigma^2\left(\frac{1}{n_D + n^{(1)}} + \frac{1}{n^{(2)}}\right)\right).$$

It is now straightforward to show that, according to (4.10), $e_{n^{(1)},n^{(2)}}$ is equal to

$$\sum_{i=1}^{K}\mathbb{E}_{m_D}\left\{\frac{\omega_{1,i}(y_{n^{(1)}})\phi\left(\frac{Y_{n^{(2)}}-E^{(1)}_{A,i}(\theta|y_{n^{(1)}})}{\sqrt{v^{(2)}_{A,i}}}\right)}{\sum_{r=1}^{K}\omega_{1,r}(y_{n^{(1)}})\phi\left(\frac{Y_{n^{(2)}}-E_{A,r}(\theta|y_{n^{(1)}})}{\sqrt{v^{(2)}_{A,r}}}\right)}\cdot\left[1-\Phi\left(\frac{\delta - E^{(2)}_{A,i}(\theta|y_{n^{(1)}}, Y_{n^{(2)}})}{\sqrt{V^{(2)}_{A,i}(\theta|y_{n^{(1)}}, Y_{n^{(2)}})}}\right)\right]\right\}. \tag{4.11}$$

This expression is essentially similar to (4.5), with updated posterior and predictive means and variances, given $y_{n^{(1)}}$. As a consequence, the criterion suggested at the end of Section 4.2.2 for the choice of the threshold η_e still holds true.

In order to illustrate the proposed methodology for SSRe we consider an application in which the normal approximation for the log hazard ratio (log HR) is used and interim analysis data are available.

Example 5: Predictive SSRe using a mixture of priors expressing opposite beliefs (B-14) In this application we consider the B-14 study (see Dignam et al. (1998), Spiegelhalter et al. (2004)) in which data from four interim analysis and final results are available. The trial was planned in order to assess a long-term protective effect of tamoxifen in preventing the recurrence of breast cancer. A sequential randomized controlled study was performed, enrolling disease-free patients after 5 years of therapy. According to the sequential design, an interim analysis was scheduled approximatively every 1-1.5 years (using O'Brien-Fleming stopping boundaries). At the beginning of the trial the planned sample size was 115 events, to detect a 40% failure reduction (corresponding to a hazard ratio of 0.6) with 85% power. Assuming a 18% event rate, this yielded a total planned sample size approximately equal to 624 patients; finally the effective number of recruited patients was 1172, because of an accrual rate lower than expected.

In Dignam et al. (1998) a Bayesian interpretation of these results is discussed under a range of prior assumptions. Using the normal approximation for the log HR estimator (see Spiegelhalter et al. (2004)), we choose here two normal priors expressing opposite beliefs, a sceptical prior $\pi_{A,1}(\theta) = N(\theta|0, 0.31)$ and an enthusiastic prior $\pi_{A,2}(\theta) = N(\theta| - 0.51, 0.31)$, where standard deviation is chosen to have

5% chance that the true difference exceeds a 40% reduction or, respectively, that a negative effect is observed ($\sigma^2 = 4$, $n_{A,1} = n_{A,2} = 41.4$). Furthermore, we center the design prior on the actual design value $\theta_D = 0.51$ (0.60 on the hazard ratio scale), with standard deviation equals to 0.19 ($\sigma^2 = 4$, $n_D = 115$). The data at the four interim analyses and the final results of the trial are summarized in Table 4.3. After each interim analysis we re-estimate the optimal additional sample size $n^{(2)*}$, needed to obtain that the predictive expectation of the probability $P(\theta < \delta | y_{n^{(1)}})$ is sufficiently large. For instance, we set $\delta = -0.22$, corresponding to a 20% reduction on the HR scale. For each interim analysis $n^{(1)} = a^{(1)} + b^{(1)}$ denotes the total number of events observed so far, with $a^{(1)}$ and $b^{(1)}$ indicating the number of events in the placebo and in the treatment arm respectively.

		$\mathbf{a^{(1)}}$	$\mathbf{b^{(1)}}$	$\mathbf{n^{(1)}}$	**logHR**	**sd**
I	after first interim	18	28	46	0.435	0.295
II	after second interim	24	43	67	0.567	0.244
III	after third interim	32	56	88	0.545	0.213
IV	after fourth interim	36	66	102	0.588	0.198
V	final results	50	85	135	0.519	0.172

Table 4.3: B14: Interim and final results on the log hazard ratio scale: $a^{(1)}$ and $b^{(1)}$ denote the number of events occurred in the placebo and in the treatment arm respectively and the total number of events is $n^{(1)} = a^{(1)} + b^{(1)}$

First of all we assign equal weights to the two prior components of the mixture $\pi_A(\theta)$ defined in (4.1). The analysis prior and the design prior are represented in Figure 4.3. After the first interim analysis, in order to reach a conclusion favouring tamoxifen, it would be necessary to observe a large number of events (for example, $n^{(2)*} = 59$, for a threshold $\eta_e = 0.75$ corresponding to the 80% of the supremum of $e_{n^{(1)},n^{(2)}}$). Moreover after each interim analysis the additional number of units required to conclude in favour of a protective effect of tamoxifen becomes larger and larger (see Figure 4.4). This is coherent with the fact that the negative results actually observed at each step, made it more and more difficult to revert the evidence against tamoxifen, to such an extent that the monitoring committee decided to stop the trial.

In the right panel of Figure 4.4 the dashed lines represent the SSRe criteria when the design prior is also updated after each interim: the evidence of the data supports a conclusion opposite to the one we expected in designing the experiment and this affects $e_{n^{(1)},n^{(2)}}$. In this case the previous threshold η_e is impractical already after the first interim, the optimal additional sample size is undetermined. If η_e is reduced to 0.44, after the first interim, we have $n^{(2)*} = 590$.

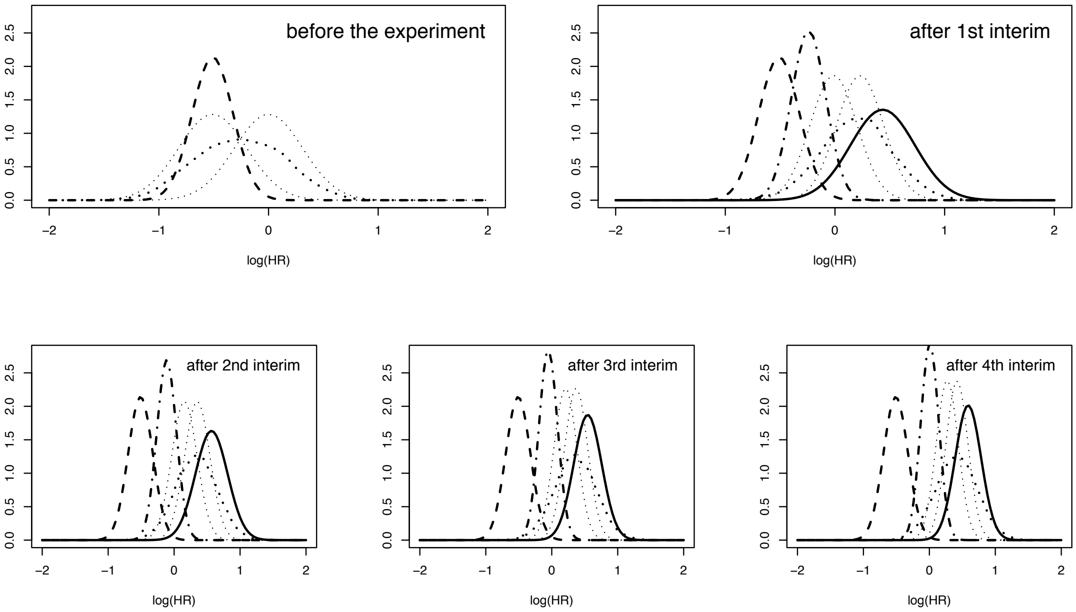


Figure 4.3: Information update at each interim point: the dotted lines represent the prior components of the mixed analysis prior, the dashed density is the fixed design prior, while the dashed-dotted curves indicate the progressive update of the design prior. The continuous line represents the likelihood at each step.

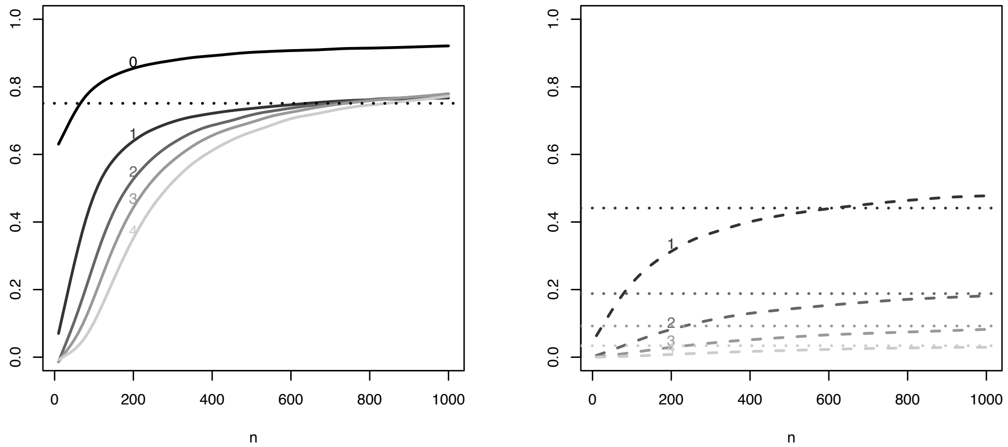


Figure 4.4: Sample size re-estimation at each interim time (denoted by the numbers from 0 to 4). Continuous lines (left panel) are referred to the case of fixed design prior; dashed lines (right panel) are referred to the case of updated design prior

		interim analysis			
	before	**I**	**II**	**III**	**IV**
$\mathbf{weight}_1$	1/2	0.87	0.94	0.95	0.96
$\mathbf{weight}_2$	1/2	0.13	0.06	0.05	0.04
$\mathbf{n^{(2)*}}$	59	638	742	732	864
$\mathbf{weight}_1$	1/3	0.77	0.88	0.90	0.92
$\mathbf{weight}_2$	2/3	0.23	0.12	0.10	0.08
$\mathbf{n^{(2)*}}$	36	543	671	714	796
$\mathbf{weight}_1$	2/3	0.93	0.97	0.97	0.98
$\mathbf{weight}_2$	1/3	0.07	0.03	0.03	0.02
$\mathbf{n^{(2)*}}$	79	800	739	787	855
$\mathbf{weight}_1$	1/10	0.43	0.62	0.66	0.72
$\mathbf{weight}_2$	9/10	0.57	0.38	0.34	0.27
$\mathbf{n^{(2)*}}$	10	503	592	669	759
$\mathbf{weight}_1$	9/10	0.98	0.99	0.99	0.996
$\mathbf{weight}_2$	1/10	0.02	0.01	0.01	0.004
$\mathbf{n^{(2)*}}$	116	799	823	826	931

Table 4.4: Optimal re-estimated sample sizes for several choices of the initial weights (weight$_1$ refers to the sceptical prior component, weight$_2$ to the enthusiastic one). Given that $e_{n_1,\infty} = 0.94$ and choosing $\beta = 0.80$, the threshold η is 0.75.

In Table 4.4 we report the optimal re-estimated sample sizes for several choices of the initial weights, with fixed design prior. The weights of the sceptical component tend to be increasingly higher, due to the evidence of the data against a protective effect of tamoxifen. This corresponds to an growing re-estimated number of required events after each interim analysis.

4.4 Concluding remarks

In summary, in this chapter we have presented a predictive methodology for sample size selection and adjustment in clinical trials, when a mixture analysis prior is used. This allows one to take into account different sources of pre-experimental information and to combine them in a simple way. Sometimes these sources actually correspond to results derived from previous studies or to opinions of several experts. It is also possible to consider "conventional" priors that reflect opposite attitudes towards the trial such as enthusiasm and scepticism. In this way we are able to incorporate a large amount of information and uncertainty on the unknown treatment effect. One of the main advantages of this approach is that it typically avoids sample size underestimation and low predictive probability of trial success.

One critical aspect of the proposed method is the choice of prior weights in the mixture. Of course, this is problem specific. However, we have discussed in the examples some strategies. In the first example (Section 4.2.4), for instance, we have compared some alternative weights assignments, such as uniform weights and weights proportional to the dimensions of the historical studies used to elicit the prior components. In the second example (Section 4.3.3) we have considered different combinations of weights for an enthusiastic and a sceptical prior and we have examined their impact on the resulting sample sizes.

The presence of several sources of prior knowledge makes it natural to plan an interim analysis and a sample size re-estimation step. This approach appears to us quite useful when available sources of prior knowledge (or experts opinions) are conflicting and when, initially, the weight of each prior in the mixture is not predominant over the others. In this case, the first portion of data allows one to adjust both the starting prior distributions, $\pi_{A,i}$ and their weights in the mixtures. Note also that, in principle, multiple sample size adjustments do not have drawbacks in a Bayesian perspective. In fact, from this point of view, repetition of the SSRe procedure just implies a sequential use of Bayes theorem. This is shown, for instance, in the example of Section 4.3.3.

This approach can be potentially applied in different situations. First of all, this methodology can be applied to other models, such as Bernoulli and survival trials. (See also Gajewski & Mayo (2006), where beta mixtures are used for non-predictive SSD). A possible extension is to consider mixtures of non–conjugate analysis priors, resorting to numerical computational methods, as discussed by Wang & Gelfand (2002).

Chapter 5

Optimal sample size for Equivalence Trials

5.1 Introduction and motivations

The first part of this thesis primarly refers to the context of superiority trials. Nevertheless, as we anticipated in Chapter 1, it is quite straightforward to adapt the proposed methodology to experimental situations with different objectives. For instance, in the present chapter we explicitly consider the case of equivalence trials, illustrating a dedicated Bayesian (robust and non–robust) approach to SSD.

An equivalence trial is designed to confirm the absence of a meaningful difference between treatments. As suggested in a recent document by the European Agency for the Evaluation of Medicinal Products (CPMP/EWP/482/99 (2000)), in this setting it is more informative to conduct the analysis by means of the calculation and examination of the confidence interval although there are closely related methods using significance test procedures (as described, for example, in Julious (2004)). It is then necessary to choose a margin of clinical equivalence by defining the largest difference that is clinically acceptable, so that a difference bigger than this would matter in practice. If the two treatments are to be declared equivalent, then the two-sided confidence interval – which defines the range of plausible differences between the two treatments – should entirely lie within the so called *range of equivalence*. This situation is schematically represented in Figure 5.1. Equivalence margins may be chosen either symmetrically or asymmetrically with respect to zero: in the following we denote the range of equivalence by $\mathcal{I} = [\theta_I, \theta_S]$. There are in practice some difficulties associated with its specification, but a detailed discussion on this point goes beyond the scope of the present work.

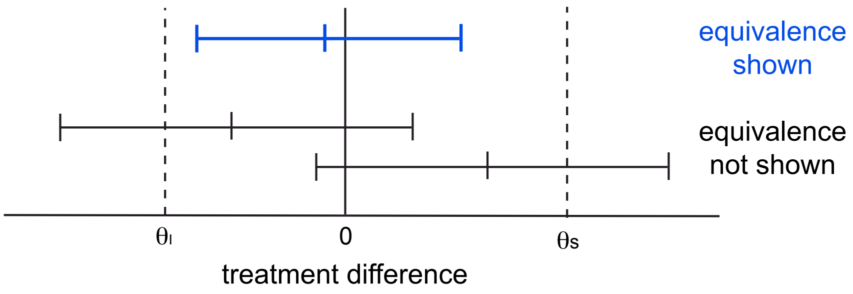


Figure 5.1: Equivalence trials

There is a large statistical literature on trials designed to establish equivalence between therapies. As stated in Spiegelhalter et al. (2004), from a Bayesian perspective it is straightforward to define a region of equivalence and calculate the posterior probability that the treatment difference lies in this range; then for example a threshold of 95% or 90% might be chosen to represent strong belief in equivalence. For further details, see for example Selwyn et al. (1981), Fluehler et al. (1983), Selwyn & Hall (1984), Breslow (1990), Grieve (1991) and Baudoin & O'Quigley (1994). A decision-theoretic formulation is proposed in Lindley (1998) and in general it can give radically different conclusions.

Bioequivalence is a slightly different problem, that is very important in practice and very popular in the literature. Two different drugs or formulations of the same drug are called bioequivalent if they are absorbed into the blood and become available at the drug action site at about the same rate and concentration (see for instance Berger & Hsu (1996)). In particular bioequivalence is of practical importance because the approval of most generic drugs in the USA and in the European Community requires the establishment of bioequivalence between the brand-name drug and the proposed generic version. This problem is theoretically interesting because it has been recognized as one for which the desired inference, instead of the usual significant difference, is practical equivalence. However in this work we focus on the generic framework of an equivalence trial, with particular reference to the aspect of SSD. In Gould (1993) a Bayesian methodology for determining the sample sizes for event rate equivalence trials is proposed. Trials for demonstrating the equivalence of active standard and test treatments generally require large sample sizes that depend on the definition of equivalence and on the overall event rate, when the outcome is incidence of an event such as mortality. The planning of sample sizes for such trials requires specification of a value for the overall event rate. This design value will often reflect the outcomes of previous trials of the standard treatment, and it is subject to uncertainty that needs some accommodation, to protect against an inadequate sample. For this reason the Author suggests to use Bayes and Empirical Bayes methods to incorporate information from one or more previous trials into the sample size calculation when equivalence means high confidence that the event rate ratio is less than some specified value.

The outline of this Chapter is as follows. In Section 5.2.2 we present a Bayesian predictive approach to sample size determination for equivalence trials. Then we deal with the problem of robustness to the prior specification (see Section 5.2.3), allowing the analysis prior to vary in a prespecified class of priors, in this case the restricted conjugate class. Finally in Section 5.3 we provide results for the normal model, illustrating some examples.

5.2 Predictive Bayesian approach

5.2.1 Preliminaries

In order to adapt the Bayesian SSD methodology of Chapter 1 to equivalence trials, first of all we need to provide a definition of success. Let us suppose that the unknown parameter θ represents a measure of comparison between two alternative treatments. As anticipated above, we consider the so-called range of equivalence $\mathcal{I} = [\theta_I, \theta_S]$, that is an interval of the parameter space with conveniently chosen bounds θ_I and θ_S, corresponding to a subset of the parameter values that indicate a negligible difference between two competing treatments. Then the experiment is considered successful if it provides evidence that $\theta \in \mathcal{I}$. Hence, we want an interval estimate of θ to be entirely included into the range of equivalence.

Let us consider a random sample $\mathbf{Y_n} = (Y_1, ..., Y_n)$ with density $f_n(y_n|\theta)$ depending on the parameter θ. We specify the analysis prior π_A and, given the observed data $\mathbf{y_n}$, we obtain the corresponding posterior $\pi_A(\cdot; \mathbf{x_n})$, as in (1.1). In the same framework introduced in Chapter 1 and recalling the objective of the trial, we actually focus on the $(1-\alpha)$-posterior credible interval for θ as a posterior quantity of interest. Hence, assuming for the sake of simplicity a unimodal posterior distribution, we have:

$$\rho_{\pi_A}(\theta|\mathbf{y_n}) = C_\alpha(\mathbf{y_n}; \pi_A) = [l_n(\mathbf{y_n}; \pi_A), u_n(\mathbf{y_n}; \pi_A)]\,,$$

where $l_n(\mathbf{y_n}; \pi_A)$ and $u_n(\mathbf{y_n}; \pi_A)$ are respectively the lower and the upper bound of the posterior credible interval. Note that $C_\alpha(\mathbf{y_n}; \pi_A)$ can be for instance a HPD interval or an equal-tail interval. Finally we can declare equivalence if $C_\alpha(\mathbf{y_n};\pi_A) \subset \mathcal{I}$, that is if its bounds simultaneously satisfy the following conditions:

$$l_n(\mathbf{y_n};\pi_A) > \theta_I \quad \text{and} \quad u_n(\mathbf{y_n};\pi_A) < \theta_S. \tag{5.1}$$

5.2.2 Criteria

It is necessary to remind once again that before the experiment, the posterior quantity of interest, that is in this case the bounds of the posterior credible interval are random quantities, denoted by $l_n(\mathbf{Y_n})$ and $u_n(\mathbf{Y_n})$ in order to underline their dependence on the random sample $\mathbf{Y_n}$. As discussed in Chapter 1 in order to account for uncertainty on the design value we use the marginal distribution of the data m_D. All we need is to adapt the SSD criteria defined in (1.5) and (1.7) to the setting of an equivalence trial. As shown in (5.1), the success of the experiment relies on two simultaneous conditions: this reflects in the definition of the following criteria, based on predictive summaries of both $l_n(\mathbf{Y_n})$ and $u_n(\mathbf{Y_n})$. In particular we have:

1. **Predictive Expectation Criterion.** Let

$$e_n^l = \mathbb{E}_{m_D}\left[l_n(Y_n)\right] \text{ and } e_n^u = \mathbb{E}_{m_D}\left[u_n(Y_n)\right] \tag{5.2}$$

   be the expected value of the bounds of $C_\alpha(\mathbf{y_n};\pi_A)$, computed with respect to the marginal m_D. The optimal sample size is then selected as the minumum n such that the expected bounds of the credible interval fall into the range of equivalence. In symbols:

$$n_e^* = \min\{n \in \mathbb{N} : e_n^l > \theta_I \ \text{ and } \ e_n^u < \theta_S\} \tag{5.3}$$

2. **Predictive Probability Criterion.** Based on the marginal m_D we define the probability that the lower bound is larger than θ_I, i.e.

$$p_n^l = \mathbb{P}_{m_D}\left[l_n(Y_n) > \theta_I\right] \tag{5.4}$$

   and, similarly, the probability that the upper bound is smaller than θ_S, i.e.

$$p_n^u = \mathbb{P}_{m_D}\left[u_n(Y_n) < \theta_S\right]. \tag{5.5}$$

   Then, given a threshold $\gamma \in (0,1)$, we select the optimal sample size as the minumum n such that these two probability are reasonably large, namely

$$n_p^*(\pi_A) = \min\{n \in \mathbb{N} : p_n^l > \gamma \ \text{ and } \ p_n^u > \gamma\}. \tag{5.6}$$

5.2.3 Robust criteria

According to the idea illustrated and discussed in Chapter 3, we derive the robust version of the predictive SSD criteria just introduced for equivalence trials. Hence, in order to define a robust version of the above SSD criteria we only need to replace π_A with a class of prior distributions Γ_A. It is the necessary to consider the *robust bounds* of the posterior credible interval as the prior π_A varies in Γ_A:

$$L_n(\mathbf{Y_n}) = \inf_{\pi_A \in \Gamma_A} l_n(\mathbf{Y_n}; \pi_A) \quad \text{and} \quad U_n(\mathbf{Y_n}) = \sup_{\pi_A \in \Gamma_A} u_n(\mathbf{Y_n}; \pi_A). \tag{5.7}$$

Therefore we say we have robust evidence that θ belongs to $\mathcal{I}$ if $L_n(\mathbf{Y_n}) > \theta_S$ and $U_n(\mathbf{Y_n}) < \theta_I$, i.e. if, for any prior $\pi_A \in \Gamma_A$, we have $C_\alpha(\mathbf{y_n}; \pi_A) \subseteq \mathcal{I}$. Then, taking into account the double condition on both the interval bounds, the following criteria are immediately given:

1. **Robust Predictive Expectation Criterion:**

$$n^*_{e,r} = \min\{n \in \mathbb{N} : e^L_n > \theta_I \;\text{ and }\; e^U_n < \theta_S\} \tag{5.8}$$

where

$$e^L_n = \mathbb{E}_{m_D}\left[L_n(Y_n)\right] \;\text{ and }\; e^U_n = \mathbb{E}_{m_D}\left[U_n(Y_n)\right] \tag{5.9}$$

2. **Robust Predictive Probability Criterion:** Given $\gamma \in (0, 1)$,

$$n^*_{p,r} = \min\{n \in \mathbb{N} : p^L_n > \gamma \; and \; p^U_n > \gamma\} \tag{5.10}$$

where

$$p^L_n = \mathbb{P}_{m_D}\left[L_n(Y_n) > \theta_I\right] \text{ and } p^U_n = \mathbb{P}_{m_D}\left[U_n(Y_n) < \theta_S\right]. \tag{5.11}$$

Of course analogous properties to those remarked in Section 3.2.2 hold true. In particular, 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. This will be illustrated in Example 6, assuming the normal model with classes of restricted conjugate priors.

5.3 Results for the normal model

Let us assume that $\mathbf{Y_n}$ is a random sample from a normal density and let us specify conjugate prior distributions for both the design and the analysis prior. Since the objective of the trial is equivalence, we need our design prior mean to assign high probability to the values belonging to the range of equivalence. For simplicity, in the following we set θ_D equal to the central value of the range (for example $\theta_D = 0$, if the range is centered on 0). On the other hand the analysis prior parameters are specified in order to model pre-experimental information on θ. Hence π_A can be centered either on negative or positive values expressing respectively scepticism and enthusiasm towards one of the competing treatments.

For example, let us suppose that a pharmaceutical company attempts to put a new drug on the market. Then the regulatory committee plans a clinical trial with the intent to show that the new drug is actually equivalent to the standard one. This yields an equivalence study with an optimistic analysis prior mean $\theta_A > 0$ and a design prior centered on 0. On the contrary, let us imagine that a pharmaceutical company wants to show that its new treatment is equivalent to a competing one, in terms of efficacy. This happens, for instance, when the company, being aware that there is not evidence enough for proving superiority, goes for equivalence. Then the new drug has chances to be approved if it guarantees some other advantages, for example in terms of safety or costs. In this case the design prior mean $\theta_D = 0$ represents the objective of the company, while the analysis prior expresses the opinion of an opponent, eventually fictitious. Note that in both situations the two-priors approach described in Section 1.2.2 allows us to formalize two different points of view about the treatments difference.

5.3.1 Criteria

Using the same results of Section 1.4, we derive the posterior distribution and the design marginal. Hence, for a given sample $\mathbf{y_n}$, the posterior bounds of the credible interval are

$$\begin{aligned} l_n(\mathbf{y_n};\pi_A) &= \frac{ny_n + n_A\theta_A}{n+n_A} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} \\ u_n(\mathbf{y_n};\pi_A) &= \frac{ny_n + n_A\theta_A}{n+n_A} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}}. \end{aligned} \tag{5.12}$$

It is then straighforward to compute the predictive quantities involved in the SSD criteria. Thus, we have respectively:

1. $$\begin{aligned} e_n^l &= \mathbb{E}_{m_D}\left[l_n(Y_n)\right] = \frac{n\theta_D + n_A\theta_A}{n+n_A} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} \\ e_n^u &= \mathbb{E}_{m_D}\left[u_n(Y_n)\right] = \frac{n\theta_D + n_A\theta_A}{n+n_A} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} \end{aligned} \tag{5.13}$$

$$2. \quad p_n^l = \mathbb{P}_{m_D}\left[l_n(Y_n) > \theta_I\right] = \mathbb{P}_{m_D}\left[\frac{ny_n + n_A\theta_A}{n+n_A} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} > \theta_I\right] =$$

$$= 1 - \Phi\left(\frac{(n_A+n)\theta_I + z_{1-\alpha/2}\sigma\sqrt{(n+n_A)} - n_A\theta_A - n\theta_D}{n\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{n_D}\right)}}\right)$$

$$p_n^u = \mathbb{P}_{m_D}\left[u_n(Y_n) < \theta_S\right] = \mathbb{P}_{m_D}\left[\frac{nY_n + n_A\theta_A}{n+n_A} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{(n+n_A)}} < \theta_S\right] =$$

$$= \Phi\left(\frac{(n_A+n)\theta_S - z_{1-\alpha/2}\sigma\sqrt{(n+n_A)} - n_A\theta_A - n\theta_D}{n\sigma\sqrt{\left(\frac{1}{n}+\frac{1}{n_D}\right)}}\right) \tag{5.14}$$

5.3.2 Robust criteria

Let us suppose now that instead of the single analysis prior π_A we want to consider a class of priors. For the sake of simplicity, we focus here on the class of restricted conjugate priors, that is defined as

$$\Gamma_{RC} = \{N(\theta|\theta_A, \sigma^2/n_A); n_A \in \left[n_A^L, n_A^U\right] \subset \mathbb{R}^+\}.$$

Under this assumption, we can exploit the results derived in Brutti & De Santis (2008) for computing the robust bounds of the credible interval in (5.7). In details, in Theorem 1 the Authors show that

$$L_n(\mathbf{y_n}) = \begin{cases} l_n(y_n; n_A^L) & y_n < \theta_A + \xi_L \\ l_n(y_n; n_A^*) & \theta_A + \xi_L < y_n < \theta_A + \xi_U \\ l_n(y_n; n_A^U) & y_n > \theta_A + \xi_U \end{cases}$$

and

$$U_n(\mathbf{y_n}) = \begin{cases} u_n(y_n; n_A^U) & y_n < \theta_A - \xi_U \\ u_n(y_n; n_A^*) & \theta_A - \xi_U < y_n < \theta_A - \xi_L \\ u_n(y_n; n_A^L) & y_n > \theta_A - \xi_L \end{cases}$$

where $\xi_k = \frac{z_{1-\alpha/2}}{2n}\sigma^2\left(n + n_A^k\right)^{1/2}$, for $k = L, U$ and $n_A^* = \frac{4n^2(y_n-\theta_A)^2}{\sigma^2 z_{1-\alpha/2}^2} - n$.

Furthermore they provide explicit expressions for e_n^L and p_n^L, using the marginal distribution m_D:

$$\begin{aligned} e_n^L &= \mathbb{E}_{m_D}(L_n(y_n)) = \\ &= l(\theta_D; n_A^L)\Phi(a_L) + l(\theta_D; n_A^U)(1-\Phi(a_U)) + \theta_A\left[\Phi(a_U)-\Phi(a_L)\right] + \\ &+ \frac{1}{\sqrt{2\pi}\lambda_m}\left[\psi_U e^{-a_U^2} - \psi_L e^{-a_L^2}\right] - \frac{z_{1-\alpha/2}^2\sigma^2}{4n}\int\limits_{\theta_A+\xi_L}^{\theta_A+\xi_U}\frac{1}{y_n-\theta_A}m_D(y_n)dy_n \end{aligned}$$

and

$$\begin{aligned} p_n^L &= \mathbb{P}_{m_D}(L_n(y_n) > \theta_I) = \\ &= \Big[\Phi(a_L) - \Phi(\sqrt{\lambda_m}(d_L-\theta_D))\Big]\cdot I_{(d_L,+\infty)}(\theta_A+\xi_L) + \\ &+ \left[\Phi(a_U)-\Phi(a_L)\right]\cdot I_{(\theta_I,+\infty)}(\theta_A) + \Big[1-\Phi(\sqrt{\lambda_m}(\max\{d_U, \theta_A+\xi_u\}-\theta_D)\Big] \end{aligned}$$

where $\lambda_m = \left(\sigma^2\left(n^{-1}+n_D^{-1}\right)\right)^{-1}$, $a_k = \sqrt{\lambda_m}(\theta_A - \theta_D + \xi_k)$, $\psi_k = \frac{n}{n+n_A^k}$ and $d_k = \theta_I + n_A^k/n(\theta_I - \theta_A) + z/n\sigma(n+n_A^k)^{1/2}$, for $k = L, U$.
It is then straightforward to derive analogous expressions for e_n^U and p_n^U:

$$\begin{aligned} e_n^U &= \mathbb{E}_{m_D}(U_n(y_n)) \\ &= l(\theta_D; n_A^U)\Phi(c_U) + l(\theta_D; n_A^L)(1-\Phi(c_L)) + \theta_A\left[\Phi(c_L)-\Phi(c_U)\right] + \\ &+ \frac{1}{\sqrt{2\pi}\lambda_m}\left[\psi_U e^{-c_L^2} - \psi_L e^{-c_U^2}\right] - \frac{3z_{1-\alpha/2}^2\sigma^2}{4n}\int\limits_{\theta_A-\xi_U}^{\theta_A-\xi_L}\frac{1}{y_n-\theta_A}m_D(y_n)dy_n \end{aligned}$$

and

$$\begin{aligned} p_n^U &= \mathbb{P}_{m_D}(U_n(y_n) < \theta_S) = \\ &= \left[\Phi(\min\{e_U, \theta_A - \xi_U\})\right] + \left[\Phi(c_L)-\Phi(c_U)\right]\cdot I_{(-\infty,\theta_S)}(\theta_A) + \\ &+ \left[\Phi(e_L)-\Phi(c_L)\right]\cdot I_{(-\infty,e_L)}(\theta_A-\xi_L). \end{aligned}$$

where $c_k = \sqrt{\lambda_m}(\theta_A - \theta_D - \xi_k)$ and $e_k = \theta_S + n_A^k/n(\theta_S - \theta_A) - z/n\sigma(n+n_A^k)^{1/2}$ for $k = L, U$.

Finally, given the above results, it is immediate to apply the robust criteria defined in (5.8) and (5.10). In the paragraph below we illustrate an application of the presented methodology.

Example 6: SSD for equivalence trials (CHART) The example considered in this paragraph is based on the CHART trial, first presented in Parmar et al. (1994) and further analysed in Parmar et al. (2001) and Spiegelhalter et al. (2004). In particular we exploit the described experimental setting to elicit the prior distributions and the necessary clinical parameters to plan an hypothetical equivalence trial, in order to draw a design scenario as likely as possible. Then we actually need to revert the point of view of the original trial whose objective was superiority (see Spiegelhalter et al. (2004)).

First of all, let us explain the general context of the CHART trial. In 1986 a new radiotherapy technique known as continuous hyperfractionated accelerated radio therapy (CHART) was introduced. The idea behind it was to give radiotherapy continuously (no weekend breaks), in many small fractions (three a day) and accelerated (the course completed in 12 days), which clearly implies considerable logistical problems. Thus, the Medical Research Council wanted to compare CHART with conventional radiotherapy in lung cancer, to assess whether CHART provides a clinically important difference in survival that compensates for any additional toxicity and problems of delivering the treatment. The results were presented in terms of hazard ratio (HR), defined as the ratio of the hazard under CHART to the hazard under standard treatment. Hence, hazard ratios less than one indicate superiority of CHART. In Spiegelhalter et al. (2004) a proportional hazards model is used, providing an approximate normal likelihood for the log HR: the estimated log HR has a normal density of mean θ and variance σ^2/m, where m is the equivalent number of events in a trial balanced in recruitment and follow-up.

In order to specify the prior distribution and the range of equivalence the opinion of expert clinicians was considered. At the beginning, the participating clinicians were enthusiastic about CHART, but there was considerable scepticism expressed by oncologists who declined to participate in the trial. Eleven opinions were elicited and Spiegelhalter et al. (2004) suggest to average the corresponding distributions, obtaining as a summary a normal prior density of mean -0.28 and standard deviation of 0.23 (corresponding to an estimated HR of 0.76 with 95% interval from 0.48 to 1.19), which implies $n_A = 74.3$. This prior could also be thought of as a posterior having observed a log-rank statistic L, such that $4L/n_A = -0.28$, and so $L = -5.5$. The expected E under the null hypothesis is $n_A/2 = 37.2$ and so the observed O under CHART is $37.2 - 5.5 = 31.7$. Thus the prior can be interpreted as being approximately equivalent to a balanced imaginary trial in which 74 deaths had occurred (32 under CHART, 42 under standard). Furthermore a sceptical prior was derived (see again Spiegelhalter et al. (2004)) with prior mean 0 and precision such that the prior probability that the true benefit exceeds the alternative hypothesis is 5%. This corresponds to a prior sample size $n_A = (1.65\sigma/\theta_A)^2 = 110$, given that $\theta_A = log(0.73) = -0.31$ and $\sigma = 2$. The eleven clinicians were also told to specify the range of equivalence, namely "a range where they felt the two regimens were approximately equivalent". The upper and lower values for the ranges were averaged and the following results were obtained. The participants would be willing to use CHART routinely if it conferred at least 13.5% improvement in 2-year survival (from a baseline of 15%), and unwilling if less than 11% improvement. Thus the range of equivalence is from 11% to 13.5%, that is on the HR scale from 0.66 to 0.71, or on the log(HR) scale from -0.41 to -0.34. The average range of equivalence is shown in Figure 5.2, with the clinical and sceptical priors derived previously.

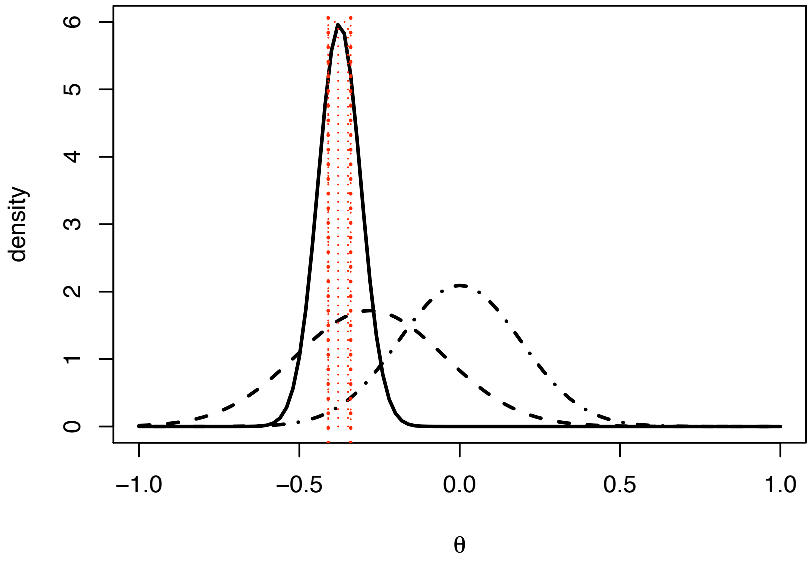


Figure 5.2: Clinical analysis prior (dashed line) with $\theta_A = -0.28$ and $n_A = 74.3$, sceptical analysis prior (dashed-dotted line) with $\theta_A = 0$ and $n_A = 110$, design prior (continuous line) with $\theta_D = -0.375$ and $n_D = 898$ and range of equivalence (dotted area) $\theta_I = -0.41$ and $\theta_S = -0.34$

Now, let us suppose we want to prove equivalence instead of superiority. In this case the above range of equivalence turns out to be too restrictive even if we choose a highly concentrated design prior on the central value of the range, for instance a normal density of mean -0.375 and standard deviation 0.067, with $n_D = 898$ (see Figure 5.2). In fact in Figure 5.3 we represent the predictive expectation of the posterior credible intervals as n increases and, adopting the SSD criterion defined in (5.3), we obtain very large values for the optimal sample size both for the clinical analysis prior (top panel) and for the sceptical analysis prior (bottom panel). Hence, we can reset the range of equivalence, in the light of the different purpose of the study. In other words, let us assume the point of view of the CHART opponents: given the logistic problems connected with CHART, the supporters of the standard treatment could consider appropriate a wider range, for instance from 5% to 15%, corresponding to $(-0.455; -0.164)$ on the log HR scale. In this case we manage to obtain much more reasonable values for the optimal sample sizes, even if we

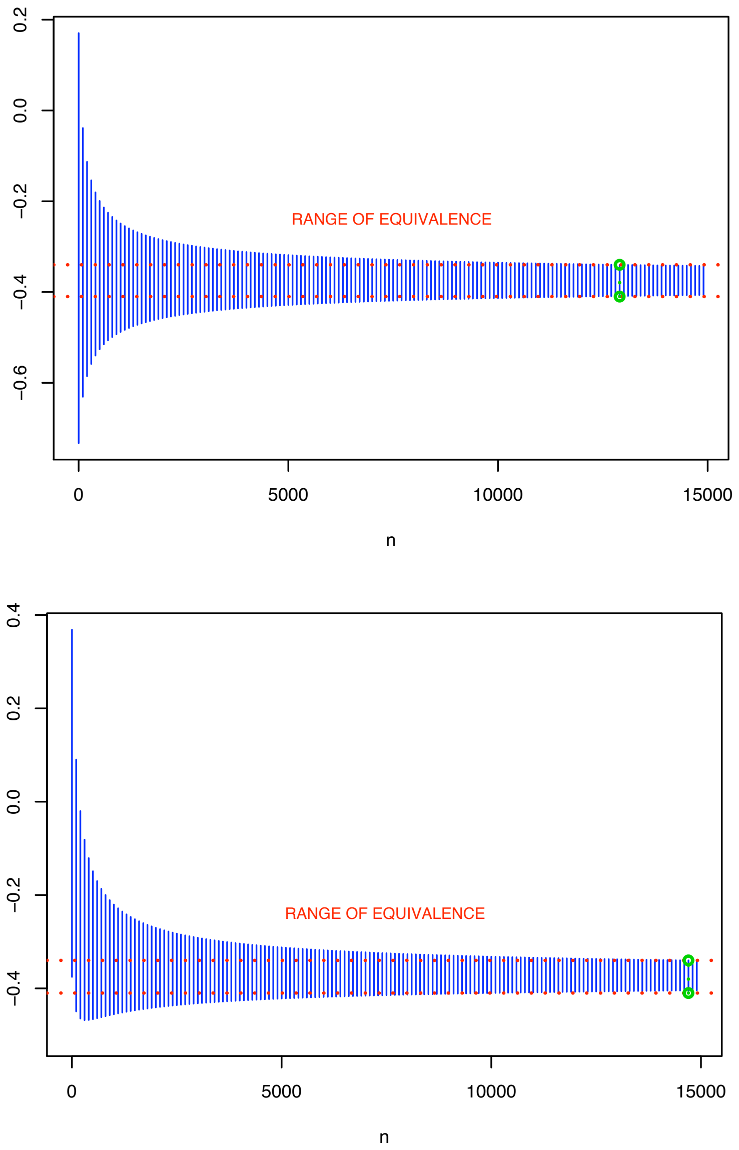


Figure 5.3: Predictive expectation of the credible interval with respect to n, assuming the clinical analysis prior (top panel) and the sceptical analysis prior (bottom panel), given the range of equivalence $[-0.41, -0.34]$ and the design prior of mean $\theta_D = -0.375$ and prior sample size $n_D = 898$. The resulting optimal sample sizes $n_e^* = 12870$ and $n_e^* = 14697$ are circled.

specify a less demanding design prior, centered in the midrange ($\theta_D = -0.3095$), and allowing for more uncertainty ($n_D = 51.9$, yielding a standard deviation of 0.278). This design setting is represented in Figure 5.4. Furthermore, in Figure 5.5 the expected range is plotted with respect to the sample size, in correspondence of the clinical analysis prior (top panel) and of the sceptical analysis prior (bottom panel): the resulting optimal sample sizes are respectively $n_e^* = 682$ and $n_e^* = 1037$.

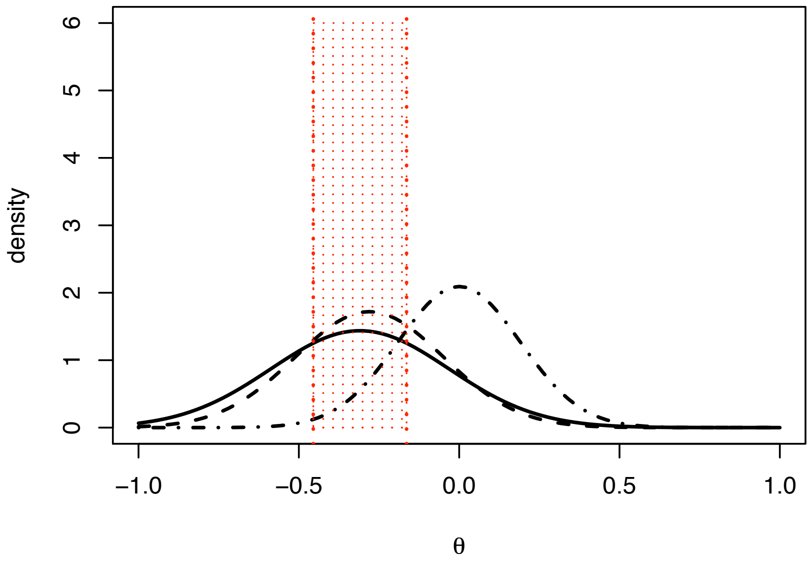


Figure 5.4: Clinical analysis prior (dashed line) with $\theta_A = -0.28$ and $n_A = 74.3$, sceptical analysis prior (dashed-dotted line) with $\theta_A = 0$ and $n_A = 110$, design prior (continuous line) with $\theta_D = -0.3095$ and $n_D = 51.9$ and range of equivalence (dotted area) $\theta_I = -0.455$ and $\theta_S = -0.164$

Similar considerations apply when we consider the predictive probability criterion defined in (5.6): the original range of equivalence actually results unpractical, while considering the range and the design parameters of Figure 5.4 we achieve a plausible value for the optimal sample size, both for the clinical and for the sceptical analysis prior (see the blue lines in Figure 5.3.2). For instance, given a threshold $\gamma = 0.5$, we have $n_e^* = 1041$ for the clinical prior and $n_e^* = 1037$ for the sceptical one. Moreover adopting the robust SSD criteria defined in (5.10) we obtain respectively $n_{e,r}^* = 1061$ choosing for instance $n_A^L = 10$ and $n_A^U = 120$ (top panel of Figure 5.3.2) and $n_{e,r}^* = 1254$ for $n_A^L = 10$ and $n_A^U = 200$ (bottom panel). The gray lines represent the probabilities that the robust bounds of the credible interval fall into the range of equivalence, as defined in (5.4) and (5.5).

Finally, in Figure 5.6 and 5.7 the gray vertical segments represent the expected robust credible intervals with respect to n for several choices of n_A^L and n_A^U. For example, using the clinical prior and a restricted conjugate class around it with $n_A^L = 30$ and $n_A^U = 100$ we obtain an optimal robust sample size of 750 observations, while the non robust optimal sample size is $n_e^* = 637$. Of course, comparing the three panel for each figure, we notice again that the wider the class the larger the corresponding optimal sample size.

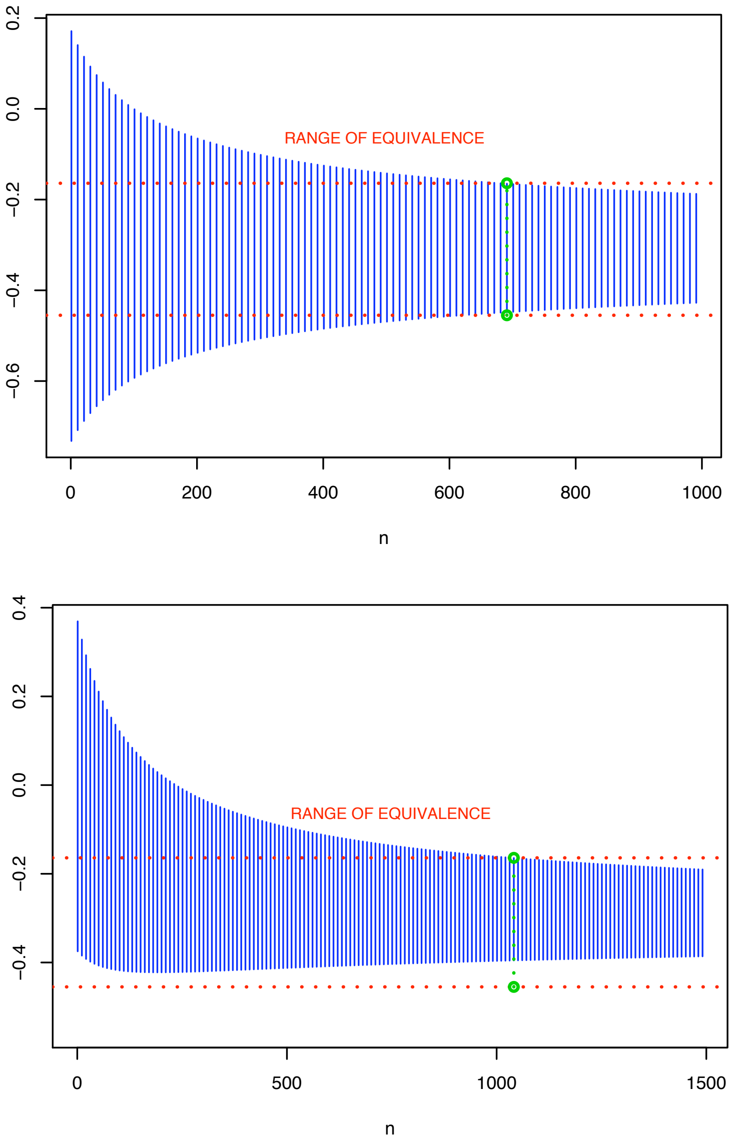


Figure 5.5: Predictive expectation of the credible interval with respect to n, assuming the clinical analysis prior (top panel) and the sceptical analysis prior (bottom panel), given the range of equivalence $[-0.455, -0.164]$ and the design prior of mean $\theta_D = -0.3095$ and prior sample size $n_D = 51.9$. The resulting optimal sample sizes $n_e^* = 682$ and $n_e^* = 1037$ are circled.

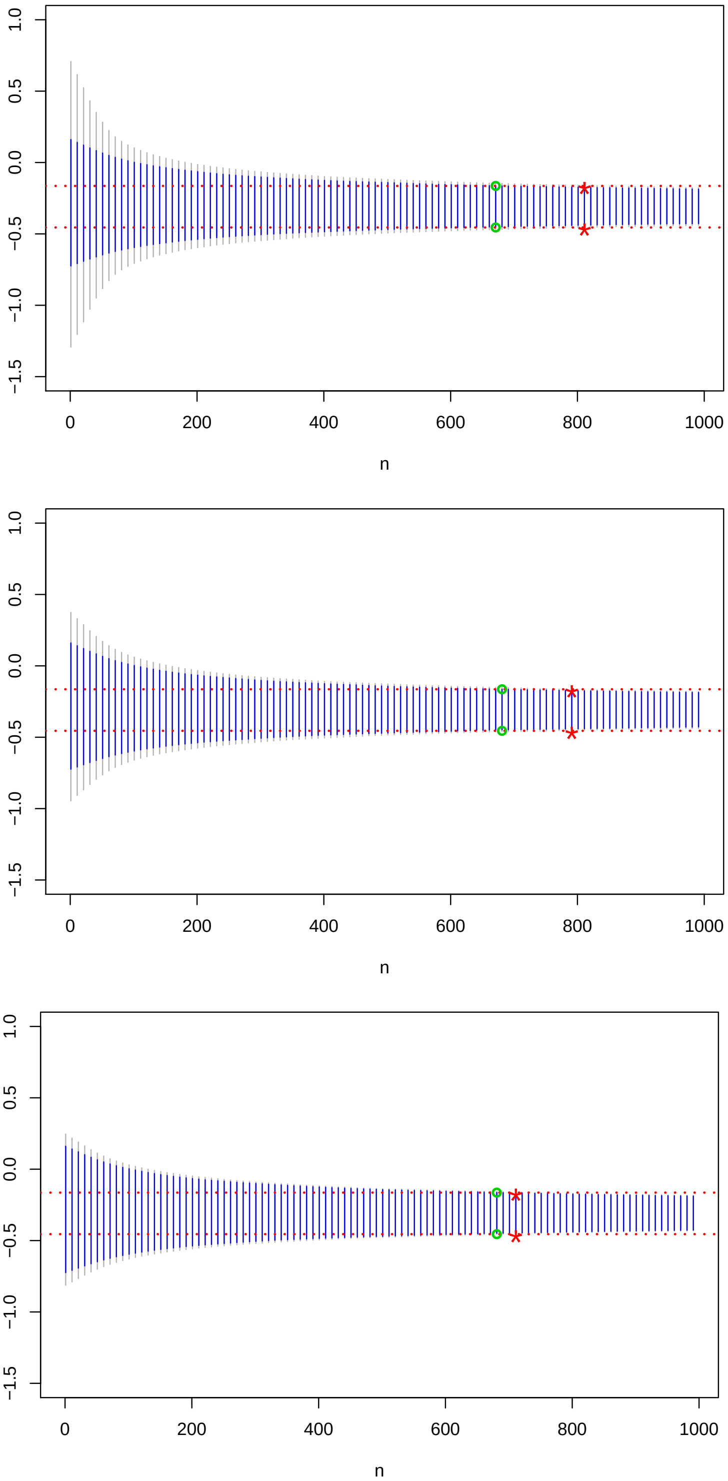


Figure 5.6: Robust and non robust SSD using the predictive expectation criterion, given the range of equivalence $[-0.455, -0.164]$ and the clinical prior. The optimal non-robust (green circle) and robust (red star) sample sizes are respecively: $n_e^* = 637$, $n_e^*, r = 801$ for $n_A^L = 10$ $n_A^U = 120$ (top panel), $n_e^*, r = 750$ for $n_A^L = 30$ $n_A^U = 100$ (center panel) and $n_e^*, r = 709$ for $n_A^L = 50$ $n_A^U = 85$ (bottom panel).

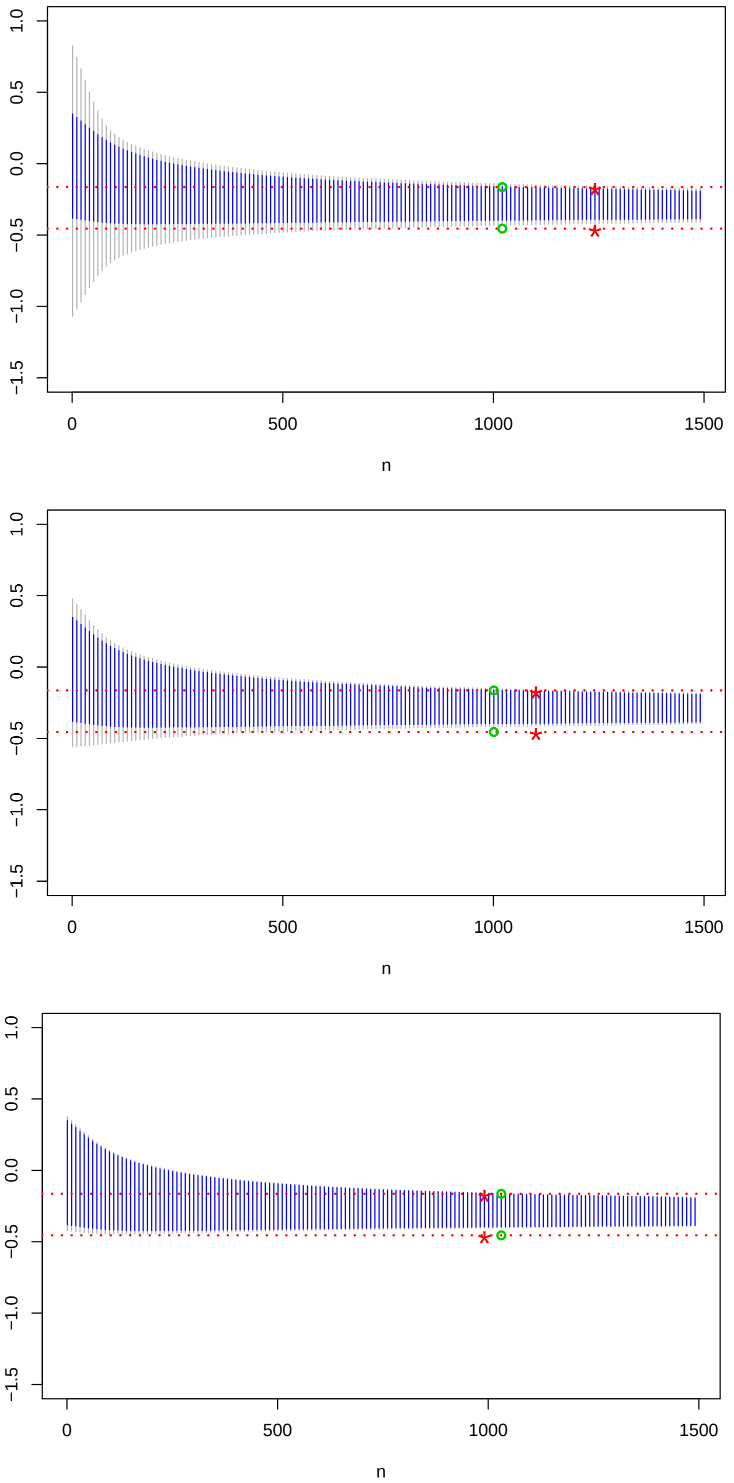


Figure 5.7: Robust and non robust SSD using the predictive expectation criterion, given the range of equivalence $[-0.455, -0.164]$ and the clinical prior. The optimal non-robust (green circle) and robust (red star) sample sizes are respecively: $n_e^* = 947$, $n_e^*, r = 1230$ for $n_A^L = 50$ $n_A^U = 150$ (top panel), $n_e^*, r = 1122$ for $n_A^L = 30$ $n_A^U = 100$ (center panel) and $n_e^*, r = 992$ for $n_A^L = 90$ $n_A^U = 120$ (bottom panel).

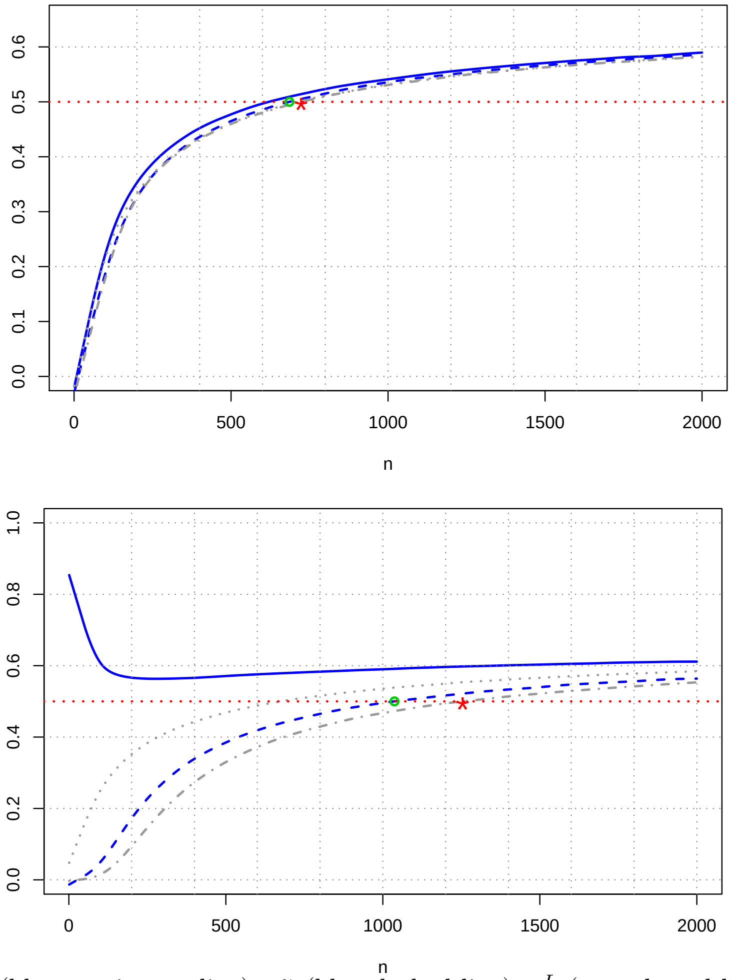


Figure 5.8: p_n^l (blue continuous line), p_n^u (blue dashed line), p_n^L (gray dotted line) and p_n^U (gray dashed-dotted line) with respect to n, given the range of equivalence $[-0.455, -0.164]$ and the design prior of parameters $(\theta_D = -0.3095, n_D = 51.9)$, for the clinical prior (top panel) and for the sceptical prior (bottom panel). The optimal non-robust (green circle) and robust (red star) sample sizes are respecively $n_e^* = 1041$, $n_{e,r}^* = 1061$ and $n_e^* = 1037$, $n_{e,r}^* = 1254$, given a threshold $\gamma = 0.5$.

5.4 Concluding remarks

In summary, in this chapter we have adapted the predictive methodology for sample size determination to equivalence trials. Thanks to the predictive approach described in Chapter 1, we are able to account for prior uncertainty and to model prior information, by specifying the design prior and the analysis prior. Specifically, due to the objective of the equivalence trials, we have pointed out that the design prior in this case should assign high probability to the values of the parameters that indicate a negligible difference between the two treatments to be compared. As for the analysis prior, according to the same idea discussed in Chapter 3, we have also addressed the issue of sensitivity to the prior specification by adopting a robust approach. Some results have been illustrated with particular reference to the normal model with the class of restricted conjugate priors, although this methodology can be potentially extended to different models and classes of priors depending on the specific context of the application.

Conclusions

In this thesis we have addressed the issue of sample size determination with special attention to the context of clinical trials. First of all, we have defined the optimal study dimension as the minimum number of observations that allows one to obtain conclusive inferential results, bearing in mind ethical considerations and budget constraints that are inevitably involved in this choice. Then, we have started noting that standard frequentist procedures for sample size calculations rely on the sampling distribution, that is a function of the unknown parameter of interest. This implies that the optimal solution heavily depends on the initial assumption on the design value for the parameter. Hence, in order to overcome this problem, we have suggested a predictive approach that enables one to model initial uncertainty on the parameter through a design prior probability distribution. This additional caution actually translates in an increased required number of units to be enrolled in the study. Moreover, we have argued that one can exploit the available pre-experimental information on the phenomenon of interest by adopting a fully Bayesian approach. Prior information can be formalized by an analysis prior distribution that in principle can be distinct from the design one. In this way, making full use of pre-experimental knowledge, we can eventually recruit a smaller number of patients. These two motivations – discussed in details in this thesis – have led us to consider a two-priors predictive approach for Bayesian sample size determination and to introduce SSD criteria based on suitable predictive summaries of a chosen posterior quantity of interest. In particular, we have derived explicit results for the normal and the binomial model and we have discussed several applications drawing on the setting of benchmark studies. This approach has been further illustrated with reference to power-based SSD methods, that are commonly used in the applications and that are shown to be a special case of the above predictive criteria.

In the second part of the thesis we have proposed some extensions of this framework, that constitute the main innovative contributions of this thesis. In particular, we have introduced a robust version of the SSD methodology by replacing a single analysis prior with a given class of distributions. We have shown the results using classes of $\varepsilon-$contaminations. Furthermore we have considered the introduced methodology in a setting in which multiple sources of prior information are available. Hence we have proposed to use as analysis prior a mixture of distributions, each formalizing the information derived from every single source. Finally, we have noted that if we want to consider clinical trials with different objectives, such as for instance showing equivalence of two competing treatments, it is possible to adapt the predictive SSD criteria to address this specific purpose. This situation is formalized in the last chapter with particular reference to the normal model and a robust methodology is also provided for the class of restricted conjugate priors.

In future research we would like to address some of the open problems in this field that certainly warrant further investigation. First of all different models and classes of prior distributions from those employed in this work could be taken into consideration. Hence the available information can be represented in the most appropriate way with respect to the context of the application. Of course, whenever it is not possible to obtain closed-form results, one can always resort to Monte Carlo approximations.

A similar methodology could also be adapted to clinical trials with multiple endpoints. In general, we can distinguish primary and secondary endpoints and we expect that one of the treatments shows a positive effect with respect to all primary endpoints. Nevertheless it is important to evaluate the impact of an innovative therapy also in terms of its potential side effects. It is then reasonable to take into account the twofold purpose of controlling both efficacy and safety, in defining the criteria for the choice of the number of patients to be recruited in the study.

A very interesting problem is the adjustment of the optimal sample size based on the data already available at a given interim analysis. This concept has already been introduced in Chapter 4, but it can also be extended up to consider a sequential approach: in practice for each enrolled patient (or cohort of patients) we have to take the decision either to stop the trial or to go on, according to a prefixed criterion. The sequential procedures have the advantage to guarantee a smaller expected number of observations with respect to the ones with prefixed sample size, other things being equal. A different dynamic is the one of two-stage designs: at the end of the first stage the experimenter has to decide, based on the observed results, whether to stop or to proceed with the second stage. This structure has to be considered in the preliminar planning of the sample size, both for the first and the eventual second stage: the Bayesian predictive approach to this problem proposed in Sambucini (2008) for binomial variables of interest, could be adapted to other settings, for instance, assuming a normal model, multiple endpoints or a robust approach.

Bibliography

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

Armitage, P., Berry, G. & Matthews, J. N. S. (2002), *Statistical methods in medical research*, (4th edn) Blackwell Science.

Baudoin, C. & O'Quigley, J. (1994), 'Symmetrical intervals and confidence intervals', *Biometrical Journal* **36**, 927–934.

Beal, S. L. (1989), 'Sample size determination for confidence intervals on the population mean and on the difference between two population means', *Biometrics* **45**, 969–77.

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

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis, 2nd edn.*, Springer.

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

Berger, J. O. & 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. & Ruggeri, F. (2000), *Bayesian robustness*, Robust Bayesian analysis, Lecture Notes in Statistics. New York: Springer-Verlag, chapter 152.

Berger, R. L. & Hsu, J. C. (1996), 'Bioequivalence trials, intersection-union tests and equivalence confidence sets', *Statistical Science* **11**(4), 283–319.

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

Bernardo, J. & Smith, A. F. M. (1994), *Bayesian Theory*, Wiley.

Breslow, N. (1990), 'Biostatistics and Bayes', *Statistical Science* **5**, 269–284.

Brutti, P. & De Santis, F. (2008), 'Avoiding the range of equivalence in clinical trials: Robust Bayesian sample size determination for credible intervals', *The Journal of Statistical Planning and Inference* **138**, 1577–1591.

Brutti, P., De Santis, F. & Gubbiotti, S. (2008*a*), 'Bayesian sample size determination and re-estimation using mixtures of prior distributions', *CLADAG* .

Brutti, P., De Santis, F. & Gubbiotti, S. (2008*b*), 'Robust Bayesian sample size determination in clinical trials', *Statistics in Medicine* **27**, 2290–2306.

Carlin, B. P. & Perez, M. E. (2000), *Robust Bayesian analysis in medical and epidemiological settings*, Robust Bayesian analysis, Lecture Notes in Statistics. New York: Springer-Verlag, chapter 152.

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

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

Clarke, B. S. & Yuan, A. (2006), 'A closed form expression for Bayesian sample sizes', *Annals of Statistics* **34**(3), 1293–1330.

Clinicaltrials.gov (2008), 'Glossary of clinical trials terms', *NIH Clinicaltrials.gov* .

CPMP/EWP/482/99 (2000), 'Committee for proprietary medicinal products (london, 27 july 2000)', *www.emea.europa.eu/pdfs/human/ewp/048299en.pdf* .

DasGupta, A. & 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**(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.

De Santis, F. & Perone Pacifico, M. (2004), 'Accounting for historical information in designing experiments: the Bayesian approach', *Annali Istituto Superiore di Sanita'* **40**(2), 173–179.

De Santis, F., Perone Pacifico, M. & Sambucini, V. (2004), 'Optimal predictive sample size for case-control studies', *Applied Statistics* **53**, 427–441.

Dignam, J. J., Bryant, J., Wieand, H. S., Fisher, B. & Wolmark, N. (1998), 'Early stopping of a clinical trial when there is evidence of no treatment benefit: Protocol b-14 of the national surgical adjuvant breast and bowel project', *Controlled Clinical Trials* **19**, 575–588.

Etzioni, R. & Kadane, J. B. (1993), 'Optimal experimental design for another's analysis', *Journal of the American Statistical Association* **88**(424), 1404–1411.

FDA (2006), 'Guidance for the use of bayesian statistics in medical device clinical trials (http://www.fda.gov/cdrh/osb/guidance/1601.html)', *Draft Guidance for Industry and FDA Staff. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health Division of Biostatistics Office of Surveillance and Biometrics* .

Fluehler, H., Grieve, A., Mandallaz, D., Mau, J. & Moser, H. (1983), 'Bayesian approach to bioequivalence assessment: an example', *Journal of Pharmaceutical Sciences* **72**, 1178–1181.

Gajewski, B. J. & Mayo, M. S. (2006), 'Bayesian sample size calculations in phase ii clinical trials using a mixture of informative priors', *Statistics in Medicine* **25**, 2554–2566.

Gould, A. (1993), 'Sample sizes for event rate equivalence trials using prior information', *Statistics in Medicine* **12**, 2009–2023.

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

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

Grieve, A. (1991), 'Evaluation of bioequivalence studies', *European Journal of Clinical Pharmacology* **40**, 201–202.

Gubbiotti, S. & De Santis, F. (2008), 'Classical and Bayesian power functions and their use in clinical trials', *Technical Report n.13, 2008- DSPSA - Sapienza Università di Roma* .

Joseph, L. & 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. & Belisle, P. (1997), 'Bayesian and mixed Bayesian/likelihood criteria for sample size determination', *Statistics in Medicine* (16), 769–781.

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

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

Lee, S. J. & Zelen, M. (2000), 'Clinical trials and sample size considerations: another perspective', *Statistical Science* **15**, 95–110.

Lindley, D. (1998), 'Decision analysis and bioequivalence trials', *Statistical Science* **13**, 136–141.

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

Marshall, R. J. (1988), 'Bayesian analysis of case-control studies', *Statistics in Medicine* **7**, 1223–1230.

M'Lan, C. E., Joseph, L. & Wolfson, D. B. (2006), 'Bayesian sample size determination for case-control studies', *Journal of the American Statistical Association* **101**(474), 760–772.

Nurminen, M. & Mutanen, P. (1987), 'Exact Bayesian analysis of two proportions', *Scandinavian Journal of Statistics* **14**, 67–77.

O'Neill, R. T. (1984), 'Sample size for estimation of the odds ratio in unmatched case-control studies', *American Journal of Epidemiology* **120**, 145–53.

Parmar, M., Griffiths, G., Spiegelhalter, D., Souhami, R., Altman, D. & van der Scheuren, E. (2001), 'Monitoring large randomised clinical trials: a new approach using Bayesian methods.', *Lancet* **358**, 375–381.

Parmar, M., Spiegelhalter, D. & Freedman, L. (1994), 'The chart trials: Bayesian design and monitoring in practice', *Statistics in Medicine* **13**, 1297–1312.

Pham-Gia, T. & Turkkan, N. (1992), 'Sample size determination in Bayesian analysis', *Statistician* **41**, 389–397.

Piccinato, L. (1996), *Metodi per le Decisioni Statistiche*, Springer.

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

Sahu, S. K. & Smith, T. M. F. (2006), 'A Bayesian method of sample size determination with practical applications', *Journal of the Royal Statistical Society. Ser. A* **17**(169), 235–253.

Sambucini, V. (2008), 'A Bayesian predictive two-stage design for phase ii clinical trials', *Statistics in Medicine* **27**, 1199–1224.

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

Selwyn, M. & Hall, N. (1984), 'On Bayesian methods for bioequivalence', *Biometrics* **40**, 1103–1108.

Selwyn, M. R., Dempster, A. & Hall, N. (1981), 'A Bayesian approach to bioequivalence for the 2 × 2 changeover design', *Biometrics* **37**, 11–21.

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

Spiegelhalter, D. J., Abrams, K. & Myles, J. P. (2004), *Bayesian approaches to clinical trials and health-care evaluation*, Wiley.

Spiegelhalter, D. J. & 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.

Spiegelhalter, D. J. & Freedman, L. S. (1988), *Bayesian approaches to clinical trials*, Bayesian Statistics 3, Oxford University Press.

Tsutakawa, R. K. (1972), 'Design of experiment for bioassay', *Journal of the American Statistical Association* **67**(339), 585–590.

Wang, F. & 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**(2), 193–208.

Wang, M. D. (2006), 'Sample size re-estimation by Bayesian prediction', *Biometrical Journal* **48**(5), 1–13.

Wasserman, L. (1992), *Recent methodological advances in robust Bayesian inference*, Bayesian Statistic 4, oxford: oxford university press edn, Lecture Notes in Statistics.

Wasserman, L. (2004), *All of Statistics*, Springer-Verlag.