Bayesian methods for sample size determination and their use in clinical trials

Stefania Gubbiotti

Abstract This paper deals with determination of a sample size that guarantees the success of a trial. We define an experiment successful if it yields a large posterior probability that an unknown parameter of interest is greater than a chosen threshold. Following a Bayesian approach, a straightforward sample size criterion is to select the minimal number of observations such that the predictive probability of a successful trial is sufficiently large. We show how this idea generalizes the traditional power-based methodology for sample study calculations. Furthermore, we address the most typical criticism to Bayesian methods – their sensitivity to prior assumptions – by proposing robust sample size criteria, that, instead of a single prior distribution, consider a class of plausible priors for the parameter of interest. An alternative way of allowing for additional uncertainty is to consider a mixture of informative prior distributions on the parameter of interest: this enables one to incorporate pre-experimental information derived from multiple sources.

Key words: Bayesian statistics, Design of experiments, Robustness, Sample size determination.

1 Introduction

Sample Size Determination (SSD), that is the choice of the optimal number of experimental units to be enrolled in a study, is one of the crucial aspects in experimental designs. The present paper is based on [16], in which we primarily refer to the context of clinical trials, although the proposed methodology can be applied to a more general experimental setting. The main purpose of a study is to observe, as efficiently as possible, the minimum number of individuals allowing inferential
analysis to be conclusive. Nevertheless, in practice one also has to deal with budget problems and ethical implications, as discussed in [18].

From a classical perspective sample sizes are usually determined either from power calculations or from formulae based on confidence interval widths (see, for example, [11]). However all standard frequentist procedures involve the use of the sampling distribution, which requires an initial guess about the true value of the unknown parameter (design value). This implies that the resulting criteria are only locally optimal and, consequently, the choice of the optimal sample size can be quite sensitive to these guessed values. This evident drawback of the standard SSD methods is one of the reasons why it is appropriate to resort to a Bayesian approach. In fact this problem can be specifically addressed by modeling prior uncertainty on the parameter values through a prior probability distribution. As highlighted by [12], assigning a prior distribution to the unknown quantities allows different plausible scenarios to be taken into consideration and local optimality to be avoided. Moreover, Bayesian methods provide a rigorous framework to incorporate in the analysis eventually available pre-experimental information – both historical information derived from previous studies and subjective opinions of experts – by specifying a prior distribution. In Berger’s words (see [3]) design problems are "naturally Bayesian": before the experiment is performed, the absence of data forces to address planning issues by using prior information. When SSD is of concern, this potentially allows us to reduce the required number of units to achieve the objectives of the trial.

In this work we propose SSD criteria based on the predictive control of a posterior quantity of interest. One of the key points is the specification of two distinct prior distributions: on the one hand the design prior models initial uncertainty on the parameter, on the other hand the analysis prior allows one to incorporate pre-experimental information on predictive posterior inference. This is the so-called two-priors approach; for an exhaustive discussion we refer to [6, 10, 11, 13, 20, 23, 24]. This approach is illustrated in Section 3 with particular reference to the power-based SSD methods. Specifically, we argue how the main drawbacks of the classical criteria may be avoided adopting the Predictive Bayesian Power function, that can be interpreted as a generalized power function.

The proposed scheme is then extended in two main directions. First of all we address the issue of sensitivity to the elicitation of a prior distribution, introducing a robust approach in Section 4. By taking into account suitable classes of prior distribution, we evaluate the impact of a specific prior on the pre-posterior analysis and, consequently, on the chosen optimal sample size. A second extension is aimed at including pre-experimental knowledge derived from several sources by combining the prior distributions relative to each source using a mixture of informative prior distributions. Furthermore, the same methodology is applied to the case of Sample Size Re-estimation (SSRe): assuming that during an ongoing trial some collected data are available at a given time point, we propose to adjust the preplanned optimal sample size, chosen at the beginning of the trial, based on the interim information.
2 Bayesian SSD: a predictive approach

The main focus of this work is the introduction of Bayesian predictive criteria for SSD based on the predictive control of the distribution of a posterior quantity of interest. As already anticipated in the Introduction, this approach serves the twofold purpose of modeling initial uncertainty with respect to the design value and of exploiting pre-experimental information on the parameter of interest, by specifying two possibly distinct prior distributions. In particular we focus on a parameter \( \theta \) and we refer to the context of Phase II and Phase III trials. A Phase II trial aims at assessing the efficacy of a new treatment, therefore \( \theta \) represents for instance a probability of success or a continuous measure of treatment effect; note that in these two cases it is straightforward to consider a binomial and a normal model respectively. Conversely, in a Phase III trial, the comparison between two treatments is of concern: \( \theta \) is a suitable measure to compare the effect of two competing therapies, such as for example the difference of proportions, the odds ratio, the hazard ratio and so on (see, among others, [19]).

Without loss of generality we restrict ourselves to superiority trials, i.e. we define the experiment successful if it yields evidence that \( \theta \) is larger than a given threshold \( \delta \). Furthermore we consider available pre-experimental information on \( \theta \) that, according to the Bayesian approach, can be formalized by specifying a prior probability distribution \( \pi_A \).

Therefore the sample \( y_n = (y_1,...,y_n) \) to be observed is a realization of the random sample \( Y_n = (Y_1,...,Y_n) \), where \( Y_i \sim f(\cdot;\theta) \) is the random variable associated to treatment efficacy. Denoting by \( f(y_n;\theta) \) the corresponding likelihood, the posterior distribution is \( \pi_A(\theta|y_n) \propto \pi_A(\theta) f(y_n;\theta) \), from Bayes theorem. Inference is actually based on a summary \( \rho_{\pi_A}(\theta|y_n) \) of this posterior distribution, that can be for instance:

a. the posterior expected value \( \rho_{\pi_A}(\theta|y_n) = E_{\pi_A}(\theta|y_n) \),

b. the posterior probability \( \rho_{\pi_A}(\theta|y_n) = P_{\pi_A}(\theta \in H|y_n) \),

where, due to the above definition of success, \( H = \{ \theta : \theta > \delta \} \) and \( \delta \) is a minimally clinical relevant threshold.

Recall now that before the experiment, being a function of the random sample, \( \rho_{\pi_A}(\theta|Y_n) \) is a random quantity as well. Hence, in order to determine the optimal sample size, we need to take into account the randomness of the data: we introduce predictive SSD criteria based on suitable summaries of the marginal predictive distribution that is obtained by averaging the sampling distribution of the data with respect to the design prior \( \pi_D \), i.e. \( m_D(y_n) = \int_\theta f(y_n;\theta) \pi_D(\theta)d\theta \). Note that \( m_D \) generalizes \( f(\cdot;\theta_D) \), that is obtained as a special case when a point-mass design prior on the single value \( \theta_D \) is chosen.

According to the choice of the predictive summary we define the following alternative criteria:

1. Predictive expectation criterion. Let

\[
e_n = E_{m_D}[\rho_{\pi_A}(\theta|Y_n)]
\]
be the expected value of $\rho_{\pi_A}(\theta|Y_n)$ with respect to $m_D$. Given a suitable threshold $\eta_e$, the chosen sample size is

$$n^*_e = \min \{ n \in \mathbb{N} : e_n > \eta_e \}.$$  \hspace{1cm} (2)

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

$$p_n = \mathbb{P}_{m_D}[A_n] = \int_{A_n} m_D(y_n) dy_n,$$  \hspace{1cm} (3)

where $\mathbb{P}_{m_D}$ is the predictive probability measure associated to $m_D$ and $A_n$ the subset of the sample space containing all the samples which yield a successful experiment at level $\gamma$: $A_n = \{ y_n : \rho_{\pi_A}(\theta|Y_n) > \gamma \}$. Given a threshold $\eta_p$ the selected sample size is

$$n^*_p = \min \{ n \in \mathbb{N} : p_n > \eta_p \} ; \quad \eta_p \in (0,1).$$  \hspace{1cm} (4)

A technical remark: Criterion 1 (that is called *effect-size criterion* by [24]) guarantees an average control on the predictive distribution of $\rho_{\pi_A}(\theta|Y_n)$, but in general a large predictive expected value does not necessarily protect one from observing small values of the posterior quantity of interest. On the contrary, as argued in [10], using Criterion 2 the sampling variability is accounted for, since one directly controls the probability of observing small values of $\rho_{\pi_A}(\theta|Y_n)$.

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 $\eta_e$ and $\eta_p$, the choice of $\delta$ and of the parameters of the design prior $\pi_D$. In [7] we propose a reasonable rule for the choice of $\eta_e$ and $\eta_p$, that is to pick these thresholds as pre-specified percentages of the maximum reachable value of the corresponding predictive quantities. This criterion guarantees that the optimization problems defined in (2) and (4) are actually well-posed. In this way we only need to specify the value of easily interpretable quantities like $\delta$, the clinically significant effect to be detected, and $\theta_D$, the true treatment effect under the design prior.

## 3 Classical and Bayesian power functions for SSD

A special case is obtained when choosing as a posterior quantity of interest the probability $P_{\pi_A}(\theta > \delta|Y_n)$ and adopting Criterion 2. It is straightforward to notice that $p_n$ actually coincides with the Predictive Bayesian Power, introduced in [22], that, as the probability of a significant result, can be interpreted as a generalized power function.

This idea is further developed in [17] by adopting the two-priors approach. In the paper we start illustrating by example some drawbacks of the classical methodology that motivate us to choose a Bayesian predictive approach, in a particularly
interesting context for the applications, such as the one of SSD methods based on the power function.

Given a test of hypotheses on a parameter $\theta$, the **Conditional Frequentist Power function** is defined as the function that associates to every single value of the parameter space the probability of rejecting the null hypothesis, conditional to the value of $\theta$. This power function clearly depends on the sample size. It is then straightforward to choose the optimal sample size as the minimum number of observations that guarantees to achieve a given power. This criterion is widely used in the applications although it has two relevant drawbacks. First of all the frequentist power and the corresponding optimal sample size noticeably depend 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 that potentially reduces the required number of subjects. Hence a double solution is needed. On the one hand it is possible to take into account the uncertainty around the design value by specifying a prior probability distribution. Nevertheless, suppose that we want the conclusions of the study to be entirely classical, namely we do not intend to incorporate prior information in the final analysis. In this way, we obtain the **Predictive Frequentist Power**, according to the so-called hybrid classical-Bayesian approach described in [22], which results in larger optimal sample sizes. On the other hand, pre-experimental information can be incorporated into the analysis by adopting the Bayesian approach. In fact the information on the unknown parameter $\theta$ can be formalized through the analysis prior probability distribution. 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. Again, it is possible either to condition with respect to an initial design value in order to obtain the **Conditional Bayesian power**, or to adopt a predictive approach to get the so-called **Predictive Bayesian power**. As for the latter we discuss in particular the possibility of specifying two distinct prior distributions with two different roles: the design prior $\pi_D$ is used for designing the experiment and in a sense it represents the goal of the trial; the analysis prior $\pi_A$ is actually used for inference and it is based on the available pre-experimental information.

As anticipated at the beginning of this section, an interesting remark is that the Predictive Bayesian power can be actually considered as a generalized power function including the other three power functions as special cases. If the design prior tends to a point-mass, i.e. a distribution that assigns probability 1 to a single point $\theta_D$, we get the Conditional Bayesian power, that is a function of the design value $\theta_D$. On the other hand if we keep a proper design prior and we let the analysis prior degenerate in a flat non-informative prior, we obtain the Predictive Frequentist power. Finally the Conditional Frequentist power comes out when we choose simultaneously a point-mass design prior and a non-informative analysis prior, meaning that both design uncertainty and prior information are ignored.
4 Robust Bayesian SSD

The elicitation of a specific prior for posterior analysis has been always a major criticism to Bayesian inference, due to its impact on the analysis and, in our specific context, on the selection of the optimal sample size. Hence, we may want to model additional uncertainty that is intrinsic to the choice of one particular prior distribution. An attempt to address this objection is represented by the robust Bayesian approach that replaces a single prior with a class of distributions providing a more flexible and realistic representation of pre-experimental knowledge. The idea is that if the range of the posterior quantity of interest, as the prior varies over the class, is small, then the differences between the various elements of the class are irrelevant and then one can use the starting prior with confidence. On the contrary, if the posterior range is not small enough, robustness is a concern and refinement of prior knowledge is needed. General principles of the robust Bayesian approach are discussed in [2, 4, 5, 26], while for specifically dedicated applications to clinical trials we refer, among others, to [9, 15].

According to this principle, a clinical trial is defined robust-successful if, for any prior in $\Gamma_A$, the chosen posterior quantity of interest $\rho_{\pi_A}(\theta|y_n)$ is larger than $\gamma$ or, equivalently, if

$$\inf_{\pi_A \in \Gamma_A} \rho_{\pi_A}(\theta|y_n) > \gamma, \quad \gamma \in (0, 1).$$

The robust version of SSD Criteria 1 and 2 defined in the previous paragraph are obtained by simply replacing $\rho_{\pi_A}(\theta|Y_n)$ with its inferior bound in (1) and (3) and using the resulting predictive summaries $e'_n$ and $p'_n$ to derive the corresponding criteria. The consequence of replacing $\pi_A$ with $\Gamma_A$ (which we assume to contain $\pi_A$), is that, in general, for any given $\delta$, $\eta_e$, $\eta_p$, and $\gamma$, the robust sample size is larger than the single-prior sample size. Similarly, for any two classes of priors $\Gamma_A$ and $\Gamma'_A$ such that $\Gamma_A \subset \Gamma'_A$, optimal sample sizes determined with the latter class are larger than those obtained with the former. In this work we consider classes of $\epsilon$-contaminated priors (studied for instance in [21]) that 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. These classes are mixtures of a base prior with classes of distributions that possess some specific features, such as for instance (i) the set of all probability distribution, which is the largest contaminating class one can consider, (ii) the class of unimodal distributions and (iii) the class of symmetric unimodal distributions. In [7, 16], we derive the bounds of the posterior quantity of interest exploiting the results by [21], for the normal model and the binomial model. Then, in order to compute the predictive summaries used in the robust SSD criteria, we resort to standard Monte Carlo approximations.
5 Using mixtures of informative priors for Sample Size Determination and Re-estimation

In this section we propose a predictive Bayesian approach to sample size determination and re-estimation in clinical trials, in the presence of multiple sources of prior information on the unknown parameter of interest \( \theta \) (see [8]).

This framework has been recently considered by [14] for Phase II clinical trials with binary endpoints. As a prior for \( \theta \), the Authors proposed a mixture of conjugate prior distributions, each representing information from a single source, with weights proportional to the degree of pre-experimental “reliability” of each source. Here we extend this approach: specifically, we consider a predictive approach for pre-posterior sample size computations, adopting the two-priors approach to SSD and we present results assuming normal endpoints.

Let \( Y_n \) be an estimator of \( \theta \), the unknown quantity of interest in a clinical trial. Let us suppose that \( K \) sources of prior knowledge are available for inference on \( \theta \), 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 \( \theta \):

\[
\pi_A, i(\theta) \quad \text{for} \quad i = 1, \ldots, K.
\]

A standard way to summarize this knowledge is to consider a mixture of the \( K \) prior distributions, that is the analysis prior distribution:

\[
\pi_A(\theta) = \sum_{i=1}^{K} \omega_0,i \pi_A, i(\theta), \quad \omega_0,i > 0, \quad \sum_{i=1}^{K} \omega_0,i = 1. \tag{5}
\]

This structure holds in the posterior distribution with conveniently updated weights

\[
\omega_1,i(y_n) = \frac{\omega_0,i m_A, i(y_n)}{\sum_{r=1}^{K} \omega_0,r m_A, r(y_n)}, \quad i = 1, \ldots, K.
\]

For the sake of simplicity we focus on the posterior probability as a quantity of interest. Hence 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 \( P_{\pi_A, i}(\theta > \delta | y_n) \) is the posterior probability that \( \theta \) exceeds \( \delta \) under prior \( \pi_A, i \), for \( i = 1, \ldots, K \).

As already shown in Section 2, we introduce a predictive summary of \( P_{\pi_A}(\theta > \delta | y_n) \) with respect to \( m_D \) in order to define a SSD criterion. For instance, adopting the predictive expectation, we compute

\[
e_n = \mathbb{E}_{m_D} \left[ \sum_{i=1}^{K} \omega_1,i(y_n) P_{\pi_A, i}(\theta > \delta | y_n) \right] = \sum_{i=1}^{K} \mathbb{E}_{m_D} \left[ \omega_1,i(y_n) P_{\pi_A, i}(\theta > \delta | y_n) \right]. \tag{6}
\]
Then Criterion 1 for the selection of the optimal sample size holds. In particular we provide explicit expressions of (6) for the normal model when a mixture of conjugate normal distributions is assumed as analysis prior.

Finally, the presence of multiple sources of prior information motivates an adjustment of the preplanned sample size after that a portion of experimental outcome has become available. Hence, we also present a SSRe methodology. In particular we follow a predictive Bayesian approach close to the one proposed by [25] that 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 above described context is that results of the interim analysis allow for an update of the weights of the components of the mixture itself. 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, \( \pi_A \), 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.

6 Concluding remarks and further developments

In this paper we present a general framework for a Bayesian predictive approach to sample size calculation, with special attention to the issues of sensitivity to the prior and of dealing with multiple sources of information. Further developments of this methodology could be possible in future research, for instance, it would be straightforward to provide an application to different models and/or different classes of prior distributions, so that the available information can be represented in a more suitable way. Of course, whenever it is not possible to derive closed-form results, one can always resort to Monte Carlo approximations, in the spirit of the simulation-based approach proposed in [24].

Another interesting application could be to clinical trials with multiple endpoints. Let us consider for example two main endpoints, efficacy and safety. In general, we expect that the experimental treatment shows a positive effect in terms of efficacy. However, it is crucial to evaluate a new therapy also based on toxicity considerations, that it may be then reasonable to take into account in defining the SSD criteria as well.

One of the appealing features of the proposed methodology is the sample size adjustment based on the intermediate results provided by an interim analysis. This mechanism can be repeated several times, yielding a sequential approach, in which, after each patient’s outcome is observed, the experimenter has to decide whether to continue or to terminate the trial according to a prefixed stopping rule. Sequential procedures have the advantage of resulting in smaller expected sample sizes with respect to the methods based on a preplanned dimension, other things being equal.
References